STATISTICAL MODELING FOR DETERMINING THE RISK FACTORS OF HEPATITIS C IN PUNJAB, PAKISTAN

BY

MUHAMMAD GHIAS

01-GCU-Ph.D-STAT-2010

DEPARTMENT OF STATISTICS

GC UNIVERSITY, LAHORE (PAKISTAN)

A thesis titled

STATISTICAL MODELING FOR DETERMINING THE RISK FACTORS OF HEPATITIS C IN PUNJAB, PAKISTAN

Submitted to the GC University, Lahore

For the award of degree of

PhD

In

STATISTICS

BY

MUHAMMAD GHIAS

01-GCU-PhD-STAT-2010

DEPARTMENT OF STATISTICS

GC UNIVERSITY, LAHORE (PAKISTAN)

DEDICATION

To

Humanity,

Hepatitis C patients

&

My Adorable Parents

DECLARTION

I, Mr. Muhammad Ghias, Registration No. 01-GCU-PhD-STAT-2010 student of Ph.D in the subject of Statistics, Session 2010-14 hereby declare that the matter printed in the thesis "Statistical Modeling for Determining the Risk Factors of Hepatitis C in Punjab, Pakistan" is my own work and has not been printed, published as research work, thesis or publication in any form, in any university, research institution etc., in Pakistan or abroad by someone else except the research papers which are attached herewith.

Dated:_____

Muhammad Ghias

RESEARCH COMPLETION CERTIFICATE

It is certified that the research work contained in this thesis titled "Statistical Modeling for Determining the Risk Factors of Hepatitis C in Punjab, Pakistan" has been carried out and completed by Mr. Muhammad Ghias, Registration No. 01-GCU-PhD-STAT-2010 for the completion of PhD degree.

Supervisor

Prof. Dr. Muhammad Khalid Pervaiz
Former Dean, Faculty of Art and Social Sciences,
Chairperson, Department of Statistics,
Government College University, Lahore

Dated: 24.02.2014

Submitted through

Mr. Jaffar Hussain
Chairperson, Department of Statistics
G.C. University, Lahore.

Controller of Examinations
G.C. University, Lahore

ACKNOWLEDGEMENT

Foremost, I am very thankful to Almighty Allah, the most beneficent and merciful, who gave me patience and mettle to complete this work. Nothing can ever be accomplished without His blessings. I can't thank Almighty Allah enough for blessing me to be part of Ummah of our beloved Holy Prophet (Pease be upon him), the source of all blessings and all knowledge.

I would like to pay my sincere gratitude to my supervisor Prof. Dr. Muhammad Khalid Pervaiz for his immense support, patient guidance, loving encouragement, and insightful discussions throughout my thesis work. His valuable instructions in developing and refining this document are acknowledged. He helped me in all respects and I feel lucky that he cared so much for my work and responded to my questions so promptly.

I am really grateful to Mr. Jaffer Hussain, Chairperson, Department of Statistics, G.C. University, Lahore for his kind support and ready cooperation during my studies. Guidance of Prof. Masood Amjad Khan, Ms. Hina Khan, Ms. Amina Shehzadi and Mr. Yasar Mahmood are also acknowledged as they always extended valuable help as and when required.

It will be unjust not to mention the contribution of Higher Education Commission (HEC) of Pakistan which granted me IRSIP scholarship enabling me to conduct a research work at University of Auckland, New Zealand, an internationally well known University. At this University, the experience, knowledge, and professional expertise gained under the supervision of Dr. Roger Marshall, Associate Professor and Dr. Simon James Thornley led to my personal development and enriched this document. I am really thankful for their memorable on-campus support which still continues. I am so thankful to Dr Zaheer-Ud-Din Babar, Head of Pharmacy Department, University of Auckland for his encouragement and valuable suggestions. I am also thankful to my friends Dr. Amtiaz Nadeem, Dr. Akhlaq Ahmed, Mr. Muhammad Usman Butt, Mr. Monis Ali Kazmi and Mr. Shagir Ahmed who provided all real support and care during my stay at New Zealand.

The help and services of all Medical Superintendents, Heads and doctors of the hospitals from which data were collected also are appreciable and highly acknowledged. I do express my feelings and thanks to Dr. Muhammad Umar, Professor of Medicine &

Gastroenterologist at Holy Family Hospital, Rawalpindi who gifted me his self written book on " Hepatitis C in Pakistan" during my visit which was quite informative for me later on.

Furthermore, I am thankful to my seniors' and class fellows Dr.Muhammad Riaz Ahmad, Dr.Abdul Qayyum and Dr. Asifa Kamal, Ms. Maryam Siddiqa and Ms. Uzma Hafeez for sharing some useful and constructive ideas.

I cannot ignore the real and meaningful cooperation of my dearest friends and colleagues Mr. Muhammad Shahid Rehman, Mr.& Mrs. Muhammad Muzzafar Khan and Mr. Faisal Javed. I am and will remain indebted to them for their overall support and cooperation in reviewing, commenting and refining this document.

I am unable to pay thanks for the kindness and loving attitude of my beloved parents who pray day and night for my success. They always support me in every hard and tough situation and never leave me alone. I am also indebted to the prayer, support and care of my brother and sisters. A special thanks to my wife whose accommodating behavior made it possible for me to complete this research. Another special thanks to my kids, Zainab Noor, Muhammad Afnan Tayyab for their support, understanding and patience towards my long work hours and for their less demanding behavior during my study.

Last but not least, I acknowledge all real support and encouragement of my all well-wishers, friends, colleagues and relatives, particularly, Dr. Muhammad Riaz Ahmed who provided his memorable support during all the way. Cooperation of Mr. Muhammad Sohail Majid in data collection and data entry is really acknowledged. Thanks to Mian Abdul Qayyum, Assistant Chief (Monitoring), P&D Department who spare me as and when needed. I cannot forget the moral support of Mr.Rehmatullah Bashir who always uplifts my affection to achieve desired goals.

Muhammad Ghias

Abstract/Summary

Hepatitis C disease is chronic in nature and has reported prevalence of 3.3% worldwide and 4.9% in Pakistan. The prevalence of HCV is high (6.7%) in the province Punjab, the province under study where about 60% of the country's total population lives. In this thesis, the study of the risk factors for infection with the virus was described. As, little attention has been given to this issue in developing countries like Pakistan. Moreover, earlier studies demonstrated that 20-40% of infected HCV patients do not report a history consistent with established risk factors. Also modes of transmission vary by region. No vaccine is currently available and treatment for complications of infection is costly. This hospital based, un-matched case-control was conducted in the largest province of Pakistan i.e. Punjab. A consecutive sample of 1,400 patients with 700 cases and 700 controls was collected. Case to control ratio was taken as 1:1.

The Punjab province is divided in 9 administrative divisions and 36 districts where Divisional Headquarter Hospitals are serving the humanity in each division. Patients of both-genders, all ages and social strata, urban-rural settings visit these hospitals for their medical treatment as the Government of the Pakistan is offering free of charge treatment. The researcher visited these 9 Divisional Headquarters Hospitals and with prior permission of hospital administration, interviewed patients through a questionnaire comprised of 56 variables. Important variables related to the socio-demographic factors, behavioral characteristics, patient's medical and family history related factors were studied. The cases and controls were identified from the inpatients, outpatients, and hepatitis clinics out of each hospital. The cases were HCV positive patients determined by routine ELISA method, while controls were negative. The data were analyzed descriptively and analytically using IBM SPSS version 19.0 and OpenEpi version 2.3.1 software at the University of Auckland, New Zealand.

The specialty of this study was the model building of the risk factors of hepatitis C infection. Although, multivariate logistic regression was a conventional as well as widely known statistical technique being implied for this purpose, however this technique exhibits some practical limitations which are handling of interaction effects and missing values etc. Alternatively, other techniques were used such as artificial neural networks (ANN) and classification trees models that have now been emerged in recent years, not so much to identify risk factors, but as prediction tools. These techniques would enable us for acquiring

much better insight towards the recognition of key risk factors of hepatitis C infection. The present analysis constitutes the first attempt to discover most pertinent risk factors of hepatitis C infection in such a comprehensive mode. This kind of analysis would certainly encourage the medical-researchers as well as statisticians who are supposed to have little or no experience with these statistical techniques.

Initially, the descriptive of each variable was given followed by univariate and multivariate analysis. It was hypothesized that the risk factors of hepatitis C differ by gender, residential area and geographic location. On these bases, therefore, different logistic regression (Gold standard) models were run on different segments of collected data to explore pertinent risk factors at their exact place. Later on, the analysis was performed with Artificial Neural Networks and Classification Trees models only on overall data and the results were compared and contrasted with Gold standard. In a multivariate logistic regression model on the overall data, 11 potential risk factors were identified. These were patient's education (OR=0.196, 95% CI:0.150-0.254); history of local migration/frequent travelling (OR=2.777 95% CI:2.005-3.846); family history of liver disease (OR=2.646, 95% CI:1.663-4.212); endoscopy (OR=2.357, 95% CI:1.340-4.147); family history of hepatitis C (OR=2.176, 95% CI:1.584-2.989); tattooing (OR=2.175, 95% CI:1.439-3.286); blood transfusion (OR=2.043, 95% CI:1.555-2.685); minor surgery by barber (OR=1.739, 95% CI:1.094-2.762); dental surgery (OR=1.661, 95% CI:1.252-2.205); major/minor surgery (OR=1.577, 95% CI:1.182-2.105) and history of injections/intravenous drips (OR=1.519, 95% CI:1.133-2.037). Of these risk factors, only patient's education had negative association with the disease status, whilst other risk factors were positively associated with hepatitis C infection. On comparing, Logistic regression model and Artificial Neural Networks Model, it was concluded that a concordant set of risk factors, from both models were identified which implies that these Models are a useful adjunctive method to identify risk factors for hepatitis C. However, despite the fact that Artificial Neural Networks model does include interaction effects in the model, it was challenging to convey the meaning of the model characteristics. Only networks can recognize these interactions and their effects are evaluated in the model intrinsically. If someone is curious to watch interaction effects, Classification Tree models are helpful which allow the discovery of pertinent factors with their multilevel interactions. In this study, Classification Trees model has also been applied which played very informative role to explore potential risk factors. This model can have explored 6 risk factors in overall data i.e. Patient's education, Local migration/travelling, Family history of Hepatitis C, No. of

persons sharing the room, Major/Minor surgery, and receipt of a blood transfusion which were strongly associated with hepatitis C infection. Importantly, these six risk factors were also identified in logistic regression as well as Artificial Neural Networks models. And even these were the most repeated risk factors in different models obtained for different settings of data. Therefore, these were referred as the most common risk factors in the region. In addition to overall data, logistic regression models were run on different segments of data like male/female settings, urban/rural settings and North/South regions of Punjab, keeping in view the differences in biological, lifestyle, behavioral and medical health-care related factors. In all these cases, it was observed that hepatitis C risk factors differ.

This study identified several risk factors for hepatitis C infection in the province Punjab by applying an in-depth analysis which have never been witnessed before, however two major sources of infection were identified as unhygienic, unsafe healthcare practices and personal behaviors & living conditions. Among these, Endoscopy, dentistry, transfusion services, surgical /gynecological procedures, minor surgery by the barbers, injected drug use and use of un-safe injections demand utmost care for not allowing hepatitis C transmission this way. Additionally, history of migration/travelling and under-shave from barber shops are also the newly identified risk factors in the region. Nosocomial infection should also be the focus of the medical professionals. This study also reveals the fact that patients living in poverty and low education were at higher risk. It is, therefore, pertinent to educate the people and disseminate awareness among them. In short, a general understanding of risk factors may guide preventive campaigns to reduce the burden of disease in the region.

Key Terms: Hepatitis C, Risk factors, Case-control, Odds Ratio, LR, ANN, CART, Punjab.

TABLE OF CONTENTS

Abbreviations

Terms	Abbreviations
ALT	Alanine Transaminase
ANN	Artificial Neural Networks
AOR	Adjusted odds ratio
AUROC	Area under the ROC Curve
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Algorithm for Detecting Interactions
CIs	Confidence Intervals
CT	Classification Trees
D & C	Dilation & Curettage
DHQs	Divisional/District Headquarter Hospitals
ELISA	The enzyme-linked immunosorbent assay
H/O	History of
HCC	Hepatocellular carcinoma
HCV	Hepatitis C Virus
HIMS	Hospital Information Management System
HL	Hosmer & Lemeshow
IBM	International Business Machines Corporation
IDUs	Intravenous drug users
ISO	International Standard Organization
IV	Intravenous
KPK	Khyber Pukhtoon Khwa
LR	Logistic Regression
MICS	Multiple Indicator Cluster Survey
MLE	Maximum likelihood estimation
MLP	Multi-Layer Perceptrons
MRI	Magnetic resonance imaging
MS	Marital Status
OPD	Out-patients Department
OR	Odds ratio
QUEST	Quick, Unbiased, Efficient Statistical Tree
RIBA	Recombinant immunoblot assay
ROC	Receiver's Operating Characteristic curve
SAS	Statistical Analysis System software
SPSS	Statistical Package for Social Sciences
SSE	Sum of Square of errors
STATA	Data Analysis and Statistical Software
STD	Sexually transmitted disease
T.T Singh	Toba Tek Singh
USA	United States of America

LIST OF TABLES

LIST OF FIGURES

List of Publications

The author published different research papers, majors are enlisted below:-

No.	Research Papers	Status
	Out of Thesis	
1	Ghias, M., Pervaiz, M. K., Thornley, S. & Marshall, R. 2012. Statistical Modelling and Analysis of Risk Factors for Hepatitis C Infection In Punjab, Pakistan. *World applied sciences journal,* 20**,** 241-252.	Impact factor 0.234
2	Ghias, M., Pervaiz, M. K., Marshal, R. & Thornely, S. 2012. Identification of Risk factors for hepatitis C in the Gujranwala district of Punjab, Pakistan. World applied sciences journal, 20, 94-101.	Impact factor 0.234
3	Ghias, M., Pervaiz, M. K. & Aslam, A. 2010. Risk Factors for Hepatitis C Virus among Urban/Rural Settings of Patients Visiting Tertiary Care Hospitals at Lahore, Pakistan. Journal of Statistics, 17, 33-46.	HEC 'Y' Cat.
4	Ghias, M. & Pervaiz, M. K. 2009. Identification of epidemiological risk factors for hepatitis c in Punjab, Pakistan. J Ayub Med Coll Abbottabad, 21, 156-161.	HEC 'Y' Cat.
	Others	
6	Ghias, M., Khawaja, K. I., Masud, F., Atiq, S. & Pervaiz, M. K. 2010. A New Approach For Estimation Of Body Mass Index Using Waist And Hip Circumference In Type 2 Diabetes Patients. J Ayub Med Coll Abbottabad, 22.	HEC 'Y' Cat.
7	Irfan, M., Nadeem, M. A., Mirza, H. G., Ghias, M., Mohsin, A. & Muttee, M. U. K. 2011. Statistical prediction model for relapse rate in chronic hepatitis C patients treated with conventional interferon and ribavirin therapy. British Journal of Medicine and Medical Research, 1, 122-131.	-
8	Khawaja, K. I., Fatima, A., Mian, S. A., Mumtaz, U., Moazzum, A., Ghias, M. & Masud, F. 2012. Glycaemic, insulin and ghrelin responses to traditional South Asian flatbreads in diabetic and healthy subjects. British Journal of Nutrition, 108, 1810-1817.	Impact factor 3.302
9	Siddiqa, M., Azad, Pervaiz, M. K., Ghias, M., Shah, G. H. & Hafeez, U. 2012. Survival Analysis Of Dialysis Patients Under Parametric And Non-Parametric Approaches. Electronic Journal of Applied Statistical Analysis, 5.	-

Chapter 1

INTRODUCTION

This Chapter entails a brief introduction of the hepatitis C disease, its epidemiology and prevalence of hepatitis C. Moreover, introduction about Pakistan and Punjab (province under study) is also given along with brief introduction of sampled hospitals and surrounding population. Study rationale and objectives are also described in this chapter including hypothesis to be tested.

1.1 Preliminary Information

Human existence is laid open to the danger of various ailments which augment the frequency and death rates in the population. Endeavors are, no doubt, being made to diminish the load of illnesses by enhancing public awareness about the etiology or cause of a disease and its pertinent risk factors. Hepatitis C infection is an emerging global health problem with a high prevalence in developing nations like Pakistan. It is a blood-borne illness brought about by hepatitis C virus (HCV) which was discovered in 1989 (Choo *et al.*, 1989).

Two third of the world's population lives in developing countries and currently about 200 million individuals (3.3%) of the world population, varying by region are infected with hepatitis C virus (W.H.O, 2003, Shepard et al., 2005). This figure is projected to increase with an additional 3-4 million people who develop new infection each year (W.H.O, 2003) and about 366,000 people dying annually from this disease (Perz *et al.*, 2006). Generally the infection is asymptomatic until complications develop, such as liver fibrosis and cirrhosis. In some instances, cirrhosis might lead to liver failure and liver cancer. Transmission of the virus generally involves blood-to-blood contact (Murphy et al., 2000a, Ho et al., 2012a, Ahmed et al., 2012, Ghias et al., 2010). Breast feeding, or sharing food, water, or towels with an affected person is considered safe (Madurga Revilla *et al.*, 2012). However in Pakistan, potential risk factors associated with hepatitis C include history of injections, dental surgery, blood transfusion, barber facial shaving and history of hospitalization (Qureshi et al., 2009, Ghias and Pervaiz, 2009c, Ahmed et al., 2012). Older, under-educated and poor people are also at higher risk (Muhammad and Jan, 2005). Before discussing the overall introduction about the disease, its prevalence and related risk factors, it is important to describe useful terminologies being used in this study. Medical terms are given below; however, statistical terminologies are introduced in Chapter No.3.

1.2 Medical Terminologies (Online Medical Dictionary)

- **Epidemiology:** Epidemiology is the study of determinants and distribution and of health-related events in a population which enables us in controlling and avoiding of health problems. An epidemiology study also involves in identifying risk factors for the disease.
- **Incidence Rate:** It is the rate of occurrence of newly identified cases in a specific population and disease.
- **Prevalence:** It explains that which percentage of a population is affected with specific disease and at a given time.
- **Risk factor:** A **risk factor** is the factor that increases a person's likelihood or susceptibility of a disease. For example, history of injections, tattooing, body piercing and blood transfusion are the known risk factors of hepatitis C disease.
- **Hepatitis:** Hepatitis is the Inflammation of the liver most probably caused by a viral infection but sometimes to toxic agents.
- **Jaundice:** A yellowish colour of the skin or body tissues may be caused by the abnormal discharge of bile pigments. This abnormal condition may be due to a disease like hepatitis A or eptospirosis and is characterized by jaundice
- **Blood borne Disease:** A disease which is transmitted through the blood or blood related products.
- **Cirrhosis:** A complication developed with chronic liver disease and often leads to jaundice, hepatic failure and ascites.
- **Fibrosis:** Fibrosis means thickness of the connective tissues most probably caused by inflammation or injury.
- **Hepatocellular Carcinoma:** It is the sever complication of liver developed by the chronic liver disease and ultimately known as liver cancer.
- **Acute disease:** This is a condition of a disease of quick onset with severe symptoms and short duration.
- **Chronic disease:** This is a condition of a disease which remains persistent or long-lasting and developed into certain complications.
- **Parenteral Route:** It is the route of transmission of disease administered through intravenous or intramuscular injections or simply occurs outside the intestine.
- **Infection:** An infection is the establishment of a pathogen to someone after invasion.
- **Sporadic:** It denotes temporal pattern of disease which occurs rarely and without regularity.

- **Nosocomial Infection:** Hospital acquired infection of the infection which is associated with hospitals.
- **Epidemic:** An infection which is affecting a large number of individuals in a community, population or region at once.
- **Morbidity:** The condition of being ill or diseased
- **Mortality:** The situation of being mortal or the proportion of deaths to population.
- **Diagnosis:** The decision of identifying a disease by observing its signs and symptoms
- **Tertiary Care Hospital:** A large or major hospital equipped with advanced diagnostic and therapeutic services which are mostly not available in general hospitals.
- **Abortion:** The termination of a pregnancy at the earliest stage due to rupture in the embryo or fetus.
- **Dilation and Curettage (D&C):** It is the treatment of surgical abortion or abnormal bleeding during the early stage of the 2nd trimester of pregnancy and simply known as D&C.
- **Cesarean Section:** It is a surgical procedure through which the wall of the mother's abdomen is cut for delivering a baby.
- **Extramarital relationship:** Adulterous or extramarital affairs**.**

1.3 Viral Hepatitis

Hepatitis means "inflammation of liver", caused by the viral infection. Till now, seven known strains of viral hepatitis have been discovered and referred as type A, B, C, D, E, F, and G. Out of these types A, B, C and E are the most prevalent. Hepatitis A and E are spread by the fecal-oral routes. Whereas types B, C and D spread through perinatal, percutaneous, blood to blood contacts, or by un-safe interaction with opposite gender (Poynard *et al.*, 2003, Bosan *et al.*, 2010). Main causes of spread of hepatitis A & E are poor sewage systems and contamination of drinking water, particularly in developing countries. These types of hepatitis become sporadic and epidemic during monsoon (rainy seasons) and flood seasons (Shahzad *et al.*, 2001, Bosan *et al.*, 2010).

Hepatitis A, E, G types are mainly divided into the acute hepatic disorder while hepatitis B, C, D known as chronic disease (Hall, 2007). When hepatitis disease lasts less than six months, it is considered as acute and chronic otherwise. Generally, hepatitis B & C lead to chronic stage and persistently cause the complications such as liver cirrhosis, hepatocellular carcinoma (HCC), hepatic insufficiency and digestive haemorrhage.

Resultantly, morbidity and mortality are increasing day by day. The situation becomes worsen in developing countries due to its inadequate diagnosis and treatment (Poynard *et al.*, 2003). However, prevention from A, B and D is possible owing to availability of their vaccines while for others vaccine is, unfortunately not available. Therefore, the dream for controlling of hepatitis B & C may come true by knowing their exact causes and implementing valid measures for virus control.

1.4 Hepatitis C Virus

Recent study is only concentrating on hepatitis C infection which is caused by the hepatitis C virus known as HCV. It is a small RNA virus with six genotypes. The virus was first time discovered in 1989 as the major causative agent of non-A, non-B hepatitis (Choo *et al.*, 1989).

This disease is known as "Silent Killers" because most of the people are carriers of this virus and they do not know about its existence in their bodies due to the fact that it is asymptomatic. Whenever, this fact reveals, complications have been multiplied. Moreover, for hepatitis C, no vaccine have becomes available so far due to antigenic drift of the virus. With no vaccine available for the foreseeable future, the burden of disease is likely to increase and Pakistan will struggle to meet the substantial health-care expense. However, very efficient treatment is now offered, which wipe out the virus in 60% of cases and minimizes progression to cirrhosis in the remaining cases. Minimum six months treatment with standard interferon and ribavirin is required to inactivate this virus in the body. For its complete treatment, rough estimate of its treatment cost is about US$700 including follow-up and investigation (Umar and Bilal, 2012). The statistics have shown that the increase in mortality has raised hepatitis C patients due to its chronic nature and HCC in most countries (El-Serag, 2002, Deuffic et al., 2002).

1.4.1 Acute & Chronic Hepatitis C

The incubation period for acute hepatitis C infection stays between 6 to 10 weeks. Data regarding acute hepatitis C is not so frequent as most of the acute patients (~80%) have no symptoms (Mast *et al.*, 1999). About 15-40% of infected patients are immune enough to clear the virus without any medication while in the acute stage of infection, whereas the other 60-85% develop chronic hepatitis C (Neumann et al., 1998, Grebely et al., 2012, Thimme et al., 2001). The patients, who become the chronic carriers of virus, develop complications like cirrhosis or hepatocellular carcinoma (HCC) complication. About 1.4 million people die

annually due to liver cirrhosis or HCC (Poynard *et al.*, 2003). In Pakistan, the lifetime incidence of hepatocellular carcinoma (HCC) is about 8% among people with established HCV infection (Tong *et al.*, 1996).

1.4.2 Symptoms of Hepatitis C Infection

As mentioned earlier, majority of HCV infected patients are asymptomatic or have nonspecific symptoms. In the acute phase of the illness, jaundice is present in less than 25 percent while other symptoms are analogous to those in other types of acute viral hepatitis, including nausea and malaise. Similarly in chronic patients, the most frequent is fatigue with subsequently other less common symptoms include nausea, myalgia, anorexia, weakness and weight loss (Lauer and Walker, 2001). ALT levels frequently fluctuate over time and may be normal or significantly rose in same patients at different periods, whereas others have persistently normal or persistently elevated ALT level (Umar and Hamama-ul-Bushra., 2006).

1.4.3 Comparison of Global Prevalence of HCV in Different Countries

Hepatitis C is of concern both to developed and developing countries and has become a global public health problem, no doubt. According to WHO estimates 2003, about 170-200 million people globally infected and this figure is projected to increase with an additional 3-4 million people who develop new infection (W.H.O, 2003, Shepard et al., 2005) and more than 366,000 deaths are attributed to HCV each year (Perz *et al.*, 2006). It has been estimated that 49.3–64.0 million adults in Australia, Asia, and Egypt are anti-HCV positive (Sievert *et al.*, 2011). The true burden is not well known in many countries because community based serologic surveys are lacking and only specific groups are targeted (e.g. blood donors, drug users, dialysis patients etc.) which are not representative of the general population (Averhoff et al., 2012, Lavanchy, 2009). Therefore, the estimate is based on weighted average for regions rather than individual countries (Alter, 2007). Moreover, variation in the study designs and testing methodology applied by different researchers may also reduce its generalizability. However, WHO statistics regarding HCV prevalence reveal that rates of HCV infection vary broadly between and within countries throughout the world **Figure 2.1**. Frank *et al*.,(2000) reported that Egypt is the leading country with highest HCV prevalence (22%). The lowest prevalence (0.01%-0.1%) has been witnessed from countries in the United Kingdom and Scandinavia (Alter, 2007). Whereas in many developed countries like Australia and most countries in Western Europe is parallel to that in the United States (<2%) (Sievert et

al., 2011, Cornberg et al., 2011). As it has mentioned earlier that HCV prevalence differ both geographically and temporally. HCV prevalence rates are higher (≥3%) in many countries in Latin America and Eastern Europe, countries of the former Soviet Union, the Middle East and some countries in Africa, and South Asia (Cornberg et al., 2011, Shepard et al., 2005, Madhava et al., 2002, Averhoff et al., 2012). On the other hand, vastly different countries such as United States, Turkey, Spain, Australia, Japan and Italy have similar average HCV prevalence (1.0%-1.9%) (Alter, 2007). However, pattern of these HCV prevalence are age-specific. This pattern illustrates that in majority of the cases HCV transmission occurred in the adult age group (20-40 years).

It is an alarming situation that Pakistani HCV serofrequency are significantly higher (4.9%) when compared with neighboring countries like Iran (0.87%), India (0.9%), Nepal (1.0%), Afghanistan (1.1%) and Indonesia (2.1%), Myanmar (2.5%), China (3.2%), Taiwan (4.4%) (Umar et al., 2010, Sievert et al., 2011, Shepard et al., 2005). However, it is believe that the epidemiology of hepatitis C in all these neighboring countries are not studied systematically due to poor sampling or study design, therefore, this comparison may not looks so realistic. In India, among the blood donors, HCV prevalence has been reported as 1.8% (Panigrahi *et al.*, 1997).

Reportedly, HCV Prevalence, among the injecting drug users (IDUs) was very high globally and about 90% of new hepatitis C infections are attributed due to injecting drug use (Hellard *et al.*, 2009). Extremely high prevalence of HCV infection have been reported in IDUs in China *i.e* 98% which is quite similar in many other regions of the world (85%–95%) (Garten *et al.*, 2004, Maher *et al.*, 2004, Page *et al.*, 2009). Another latest study from Taiwan reveals that seroprevalence rate of HCV was 74.4% among the Heroin users (Ottenbacher *et al.*, 2004).

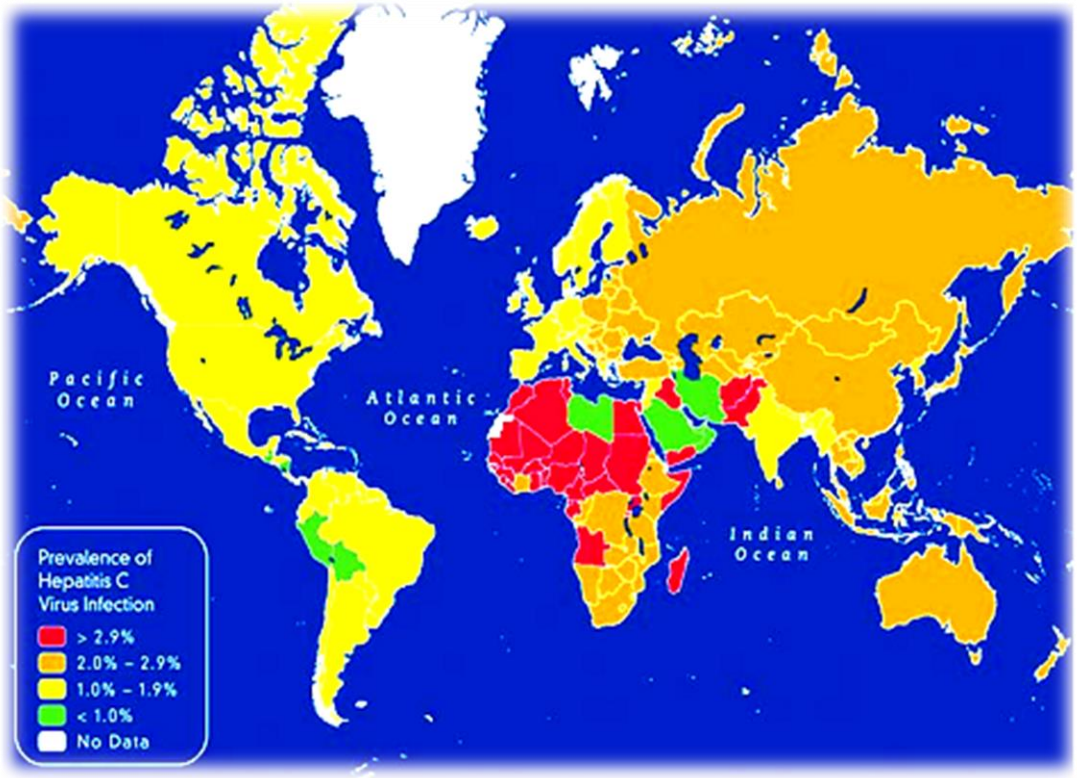


Figure 2.1: Global prevalence of chronic HCV(reprinted from (Holmberg, 2012).

1.4.4 Comparison of Risk Factors and Modes of Transmission

Most frequent way of transmission of hepatitis C infection is the parenteral route which means that infection may be transmitted through unscreened blood transfusion, exposure to infected blood and its products; injecting drug use, reuse of syringes, tattooing, organ transplantation, body piercing, health-care exposure, sharing personal items (razors, toothbrush and nail cutters). However, this disease may also transmit through the non-parenteral route or perinatal transmission. This includes sexual and household contacts (Alter, 1995, Umar and Bilal, 2012, Alavian, 2010). The parenteral transmission is considered the most efficient and best recognized mode of HCV acquisition which accounts for 30%-60% of HCV cases (Alter *et al.*, 1999).

From studies, numerous risk factors have been reported which are associated with HCV infection. The most common parenteral risk factors are generally injection therapy and blood transfusion (Jafri and Subhan, 2010) in developing countries. It is estimated that at least 12 billion syringes are sold every year and many of these are un-sterilized which can cause HCV (Kermode, 2004) mostly, in low income countries. Pakistan is the leading country where un-necessary therapeutic injections are frequently administered and an average of 13.6

injections are received every year per person (Janjua et al., 2006b). Other aspect of frequent use of injections is associated with the injecting drugs users (IDU) and about 90% of newly identified hepatitis C infection is due to the injection drug users world-wide (Hellard *et al.*, 2009). A study from United States reported that the strongest risk factor of HCV infection is injecting drug use (IDU) which increases the risk by 149 folds in those individuals who are consistent with history of injecting drug use (Armstrong *et al.*, 2006). While a recent study in Baluchistan, Pakistan reveal an associated risk of hepatitis C due to IDUs as 29.95 (95%CI: 7.06–127.02) (Ahmed *et al.*, 2012). In Pakistan, about 5 million are the drug users, of which 15% were regular IDUs (Control *et al.*, 2003)

Other major risk factor is the use of unsafe, un-screened Blood Transfusion (Janjua et al., 2010, Qureshi et al., 2009, Shazi and Abbas, 2006). In developed countries the risk of transmission of HCV infection through blood transfusion has been reduced to zero. However, in developing countries like Pakistan still it is a contributing factor due to the paucity of awareness and health care system. In Pakistan, the health care system is of a multipronged base, where different personals e.g. quack, homeopaths and untrained health providers are playing their major role in spreading out this infection. This situation is severing in rural areas where 66% of the Pakistani's population inherent with low education and poor socioeconomic status. WHO figure reveal that about 1.2 to 1.5 million transfusions are made annually in Pakistan. A study by Luby *et al*.,(2006) demonstrated that about 50% blood transfusions at Karachi blood banks take on by the paid blood donors where practices fell below WHO standards.

Transmission of HCV infection through dental treatment/surgery is another potential risk factor which is a cause of poor standards of hygiene and cleaning of dental instruments (Janjua et al., 2010, Coates et al., 2000). In Pakistan, about 10%-60% of the patients have a dental treatment history in multiple times (Umar and Hamama-ul-Bushra., 2006).

Interfamilial and perinatal transmission of HCV infection is also documented but the risk of transmission is petty low (de Waure *et al.*, 2010). However, there are number of questions that yet to be answered. Two imperative groups' *i.e* household contacts and inter-spousal exposure are considerable sources. A study from Egypt reported that the probability of wife to husband and husband to wife transmission was estimated to be 34% and 3% respectively (Magder *et al.*, 2005). This study also illustrated that about 6% of married individuals received infection from their spouses. Thus the risk of transmission from wife to husband is significantly different than husband to wife. Another study from central Africa reported that HCV antibody was found in (6.7%) household contacts, including (35.7%)

partners and (7%) relatives (Ndong-Atome *et al.*, 2009). A meta-analysis of 25 studies in Italy reveal that the highest pooled prevalence among partners of cases was 14.7% (de Waure *et al.*, 2010). Thus, a HCV patient household can be a risk factor for HCV, therefore, counseling of these households should be mandatory (de Waure *et al.*, 2010). Transmission of hepatitis C can also occur among family members other than spouses e.g. parents, brothers, sisters, children etc (Magder *et al.*, 2005) but highest incidence rate has been observed in children <10 years living in household with HCV positive parents (Mohamed *et al.*, 2005). In this perspective very few studies are available from Pakistan, one by Irfan and Arfeen (2004) from Islamabad (Capital city), giving prevalence of 4.34% between the spouses. While another study showed that positivity of HCV is expected in 38% spouses of index cases (Qureshi *et al.*, 2007).

Other important risk factors prevailing in Pakistani population are surgical procedures, barber shaves, sharing of razor, history of hospitalization, family history of hepatitis C, body piercing, and needle click etc.,(Ahmed et al., 2010b, Alavian and Aalaei-Andabili, 2011, Ali et al., 2009b, Bari et al., 2001, Ghias and Pervaiz, 2009c, Ahmed et al., 2010a). Amongst the socio-demographic factors; poor socio-economic status and low education are major associates.

Additional details regarding HCV disease have been added from various other studies along with their findings methodology and statistical analyses in forthcoming literature review Chapter-2.

1.4.5 Prevalence of HCV Infection in Pakistan

Prevalence of HCV in Pakistan is among the highest in the world and showed that about 10 million people are living with HCV, with significant morbidity and mortality (Waheed et al., 2009, Ali et al., 2009b). Unfortunately, Pakistan has no national reporting system, therefore, the epidemiological studies are sparse due to the high cost of HCV testing (Qidwai *et al.*, 2010). However, only few studies pertaining to prevalence of HCV infection have been reported in literature which showed that prevalence of HCV infection differ by regions or settings of patients. For example, in Pakistan, the seroprevalence of HCV in pediatrics was 2.1% (range 0.4–5.4%), blood donors (2.8%), adult general population (4.9%), pregnant women (5.3%), healthcare workers (5.5%) and other high risk group include household contacts of HCV infected patients (19.0%), multi-transfused population of patients with Thalassemia or Hemophilia (47.2%), patients undergoing chronic dialysis (38%) (PMRC, Ali et al., 2009b). On the other hand Waheed *et al.*, (2009) presented a thorough

systematic review of hepatitis C in Pakistan which includes about 91 different studies dating from 1994 to May 2009 and the %age prevalence of HCV was found between 4.95% ± 0.53% in the general adult population whereas, very high prevalence was observed in IDUs i.e. 57%. Recently, first population based survey on the actual prevalence of hepatitis B and Hepatitis C in Pakistan was carried out by the "Pakistan Medical Research Council" which manifests that overall prevalence of hepatitis C in Pakistan is 4.9% but relatively high figure found in province Punjab (6.7%) (PMRC) with wide variations in different areas within province Punjab (range 0.4–31.9%), the Province which is under study. A recent study from two twin cities (Rawalpindi & Islamabad) reveal that an unusually high prevalence of HCV in Rawalpindi (17%) vs. Islamabad (4%) (Satti *et al.*, 2012).

Epidemiologically, worldwide pattern of HCV infection has been categorized into three prevalence zones. High prevalence zone ranges between 8% to 12%, Intermediate prevalence zone (2%-8%) and low prevalence zone (<2%). Pakistan lies in the intermediate zone but there is a significant variability of predominance of HCV infection has been reported in general population among four provinces (Umar and Hamama-ul-Bushra., 2006). For example, a community based study was conducted at Hafizabad, reported the prevalence (6.5%) (Luby *et al.*, 1997). Similarly, another study by Aslam and Aslam, (2001) was carried out in selected groups in the Punjab regions of Pakistan and found that reported prevalence of 16% in Lahore whilst a much higher (23.8%) in the Gujranwala district, 20.6% Faisalabad (Ahmad *et al.*, 2007). In this study, it was further noticed that HCV prevalence increases with an increase in age and children have a prevalence of 1.3%. In the same way, another study among three union councils in Mansehra discloses the overall prevalence of HCV (10.3%) in the region (Jamil *et al.*, 2010).

This prevalence rate is different in other three provinces of Pakistan; Sindh (5%), Baluchistan (1.5%) and in Khyber Pakhtunkhwa (1.1%) (Umar and Bilal, 2012). While another latest study from Kech district, Baluchistan revealed the figure 5.5% (95% CI: 4.5%-6.5%) (Ahmed *et al.*, 2012) with an addition of higher rate in male. Even ethnic variations were also reported between the Punjabi, Pathan, Kashmiri and Afghan refugee.

1.5 Epidemiological Studies

Different studies were referred from different countries in subsections 1.4.3, 1.4.4 and 1.4.5 which give brief illustration of global prevalence of HCV infection, its epidemiology in world and Pakistan's perspectives. This may helps in understanding that burden of disease

and its routes of transmission are country-specific. In comparison with the United States and other developed countries, prevalence of HCV in developing countries is substantially high which demands for good, comprehensive hepatitis C controlling program on urgent basis. However, mode of transmission is fairly well defined and often involves with blood to blood contact, exposure to un-saved syringes and un-screened blood transfusion although the causes by which these exposures occur are country-specific (Averhoff *et al.*, 2012). Different studies have shown that risk factors of HCV infection are also different in developed and developing countries. The most prominent risk factor associated with HCV infection in developed countries is injection drug use. While in developing country, injection history, dental and general surgery, sharing of needles and tooth brushes etc are the prominent risk factors of hepatitis C. Therefore, it becomes imperative to identify country-specific routs of transmission for its prevention which may also varies by population and level of economic development (Averhoff *et al.*, 2012). For this purpose, the role of research work cannot be ignored which really help us to establish different strategies for controlling and avoiding on scientific grounds. These types of studies also help to curtail the incidence of disease by enhancing general public awareness and education.

As it has been mentioned earlier that Egypt is confronting with largest burden of HCV infection in the world. The main cause for such a high prevalence in Egypt was the mass treatment campaign of schistosomiasis patients using poorly sterilized glass syringes (Darwish *et al.*, 1993). Previous studies also showed that other causes of infection in Egypt were the surgical operations and dental procedures due to poor medical care standards (Habib *et al.*, 2001). It was the research which explored the causes of disease spread and helped in adopting the measures for amelioration of health conditions in Egypt. For example, the annual incident of hepatitis C infection among dialysis patients was reduced from 28% to 6% during 2001-2006. However, a latest study from Egypt manifests that, in general population, no statistical decline has been observed so far in HCV prevalence in Egypt (Mohamoud *et al.*, 2013). Similarly, In China, Sun *et al.*, (1991) first time reported in 1990 that plasmapheresis donation was the main cause of HCV infection in the Hebei Province. Accordingly, all blood products have been controlled stringently in China due to which no such event has been reported in this Province. Moreover, a recent study from China explored that blood transfusion, injecting drug use and interaction with opposite gender were the significant risk factors (Zhao *et al.*, 2013). The studies from Iran also has demonstrated that the prevalence of

HCV infection among patients on hemodialysis in the whole country has decreased from 14.4% in 1999 to 4.5% in 2005 (Alavian, 2007, Hosseini-Moghaddam et al., 2006).

It is also worth mentioning, although various corrective measures have been taken against the disease however, considerable results have not been witnessed due to one reason or the other. It implies that not only research work be carried out, but also effective implementation mechanism must be adopted in the light of valuable researches.

1.6 Introduction about Pakistan

This study has been conducted from the Punjab which is the largest Province of Pakistan. A brief introduction about Pakistan as well as province of Punjab has been submitted to explain social, cultural and geographical aspects. Moreover, comparison of education and health facilities has also been discussed.

Pakistan is a sovereign nation in South Asia subcontinent, officially designated as the Islamic Republic of Pakistan. It got independence from the British India on 14th August, 1947. At present, its population is more than 180 million with a growth rate of 1.6% (Wing, 2011, book, 2012). It has emerged as 6th most populous country in the world. The majority of Southern Pakistan's population lives along the Indus River. On religion basis, about 97% population belong to religion Islam leaving 2.31% Christian and Hindu minorities (0.69%). About 64% inhabitants belong to rural areas.

Pakistan occupies a truly significant geostrategic location amongst the important regions of South Asia, Central Asia and Western Asia that are shown in the given map **Figure 2.1**. The total area of the country is 796,096 km^2. It is bordered by India on the East, Iran on the West, Tajikistan in the North, China on the Northeast, Afghanistan on the Northwest.

Pakistan is a federation of four provinces *i.e* Punjab, Sindh, Baluchistan, Khyber Pukhtoon Khwa with an addition of capital city (Islamabad) and a group of Federally Administered Tribal Areas (FATA) in the Northwest. The administration of Pakistan practices genuine purview over the western parts of the questioned Kashmir region which has organized into the separate political entities Azad Kashmir and Gilgit–Baltistan. The Gilgit–Baltistan Empowerment in 2009 by giving it self-government and assigned a province-like status. Each province embraces different social and economic status. Pakistan is an ethnically diverse and multilingual country with more than sixty languages that are spoken around the country. National language of Pakistan is Urdu which is understandable by the 75% of

Pakistanis. On the other hand, official language is English which is used in government offices and courts. Punjabi & Saraiki are the common languages of Punjab. Similarly, Sindhi, Pashto and Balochi are the provincial languages of respective provinces *i.e* Sindh, Khyber Pakhtoon Khwa and Baluchistan.

Geographically Baluchistan is the largest province whilst Population wise Punjab is the biggest one comprising 65% of the country's total population. The country's largest cities are Karachi (metro area), Lahore, Faisalabad, Rawalpindi and Gujranwala.

The overall literacy rate in the country is 58.5%; male (70.2%) and female (46.35%). However, this literacy rate absolutely varies from region to region. For example, in tribal areas, female literacy rate is 3%. Majority of the people have awareness issues about the common infectious diseases despite the fact that they are literate. Pakistan is a developing country with rapid economic growth and is ranked at 27th largest country in the world by purchasing power. The economy is semi-developed wherein agriculture accounts for 21.2% of the GDP. As per Economic Survey 2011-12, real GDP growths was estimated as 3.7% (Wing, 2011). Regarding health and education, Pakistan is considered lowest among the nuclear power countries which are quite alarming. Only a little share for public healthcare services is being allocated since last years. In 2010 about 2.2% of GDP was allocated for this purpose. It is said that terrorism is the main hindrance to utilize the recourses at their full potential. According to the recent estimates, about 20% of the population is living below the international poverty line i.e. US$1.25 per day. Thus, majority of the people belong to poor families and live in rural areas (65%) of Pakistan which tend to be considerably more confronting and awful condition with regard to health and wellbeing services. A brief discussion on health facilities in Pakistan and Punjab province are given in the forthcoming section. Pakistan occupies highest number of illiterates in the world. The situation becomes worsen for women and rural people (Rahman and Uddin, 2009).

The following maps of Pakistan and Punjab are constructed by using the software (Pak-info) version 6.1.

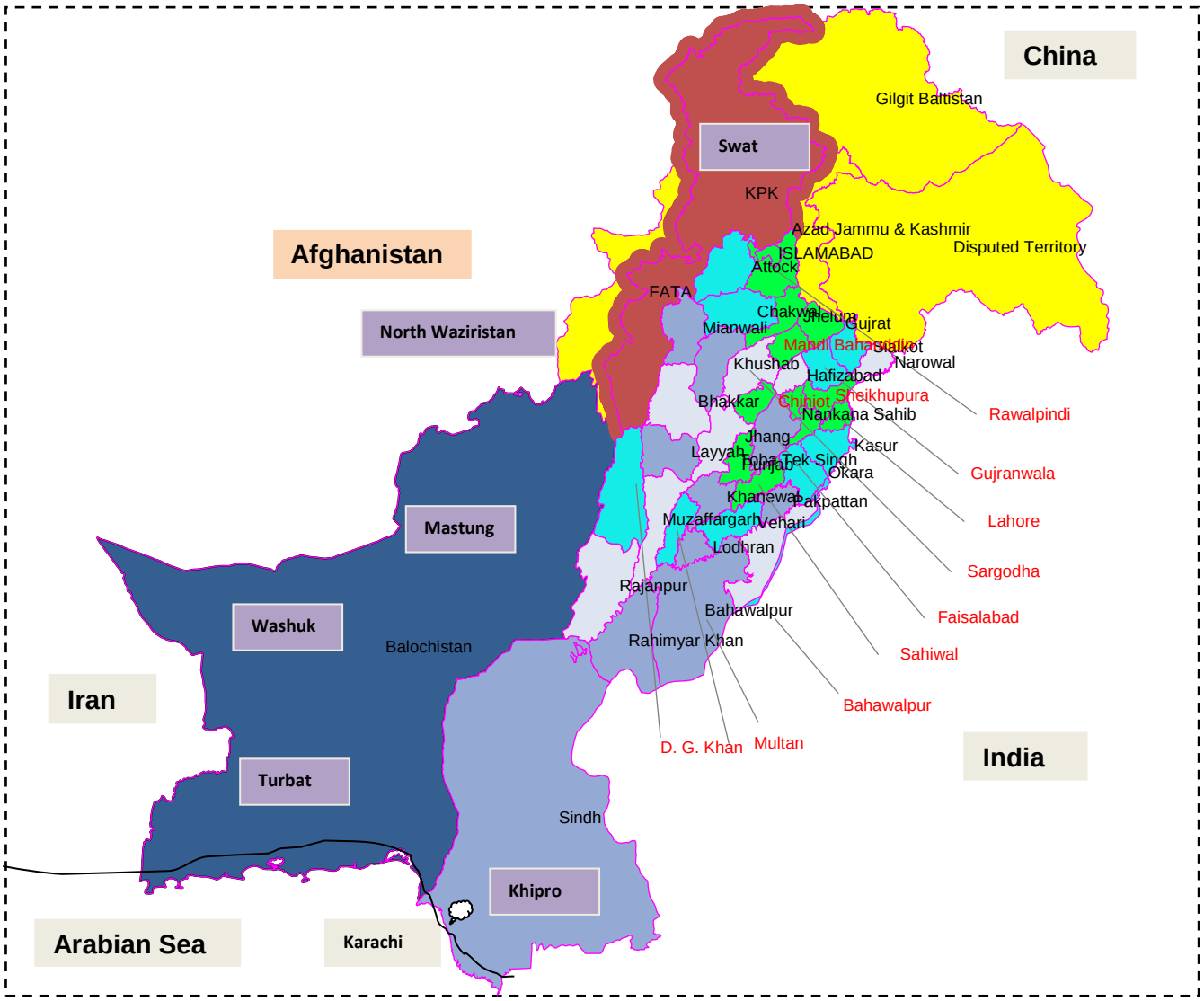


Figure 2.2: Map of Pakistan

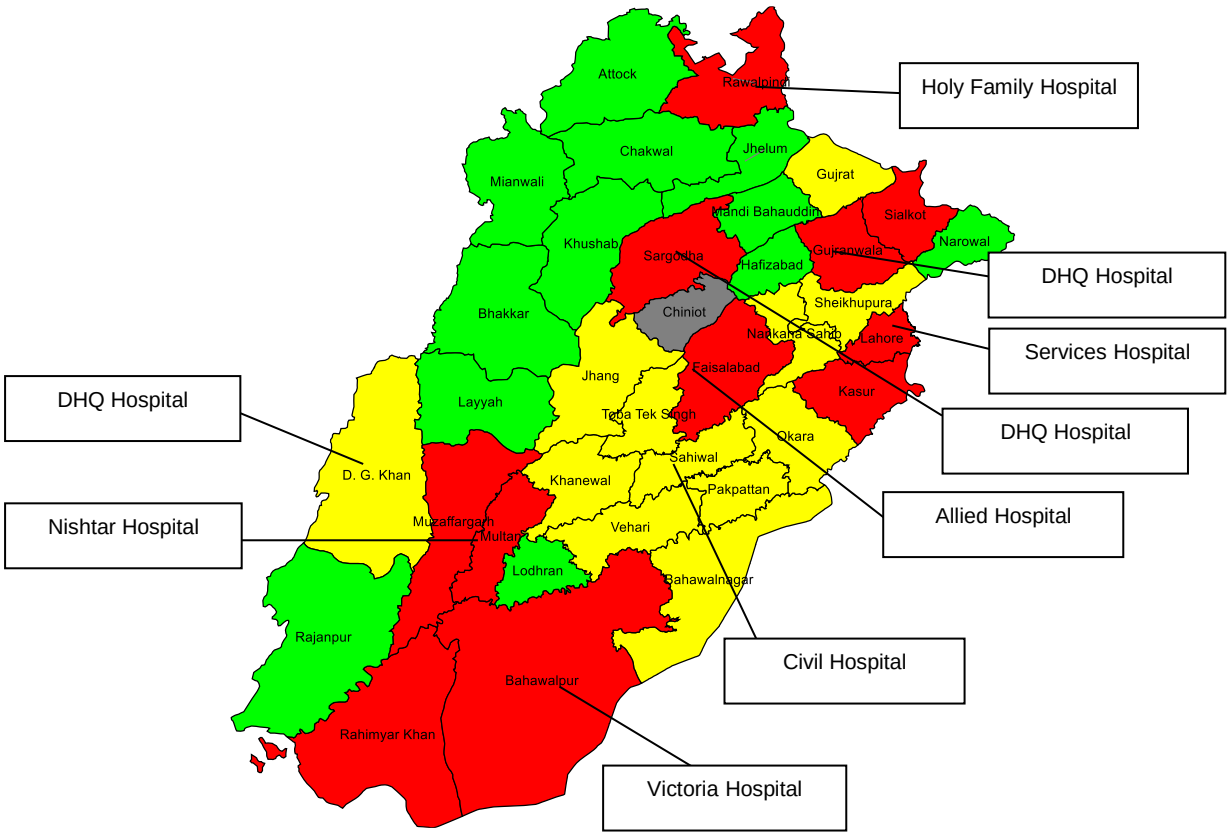


Figure 2.3: Map of Punjab Province

1.6.1 Punjab (Province under Study)

The word PUNJAB is derived from the Persian word Punj (five) and Ab (water) referring to five rivers that flow through the province. These rivers are Jhelum, Chenab, Ravi, Sutlej and Beas. Punjab is the most developed and thriving province of Pakistan. It is biggest province of the Pakistan comprising of about 60% of the country's total population. In 2012 total population was estimated 94.40 million as of mid 2011 with population density of 450/km2 (1,200/sq mi). The majority of population belongs to Islam with minimum presence of Christian minority. The temperature is generally hot, with distinct variations between summer and winter. Temperature lies between -2°C to 45°C, but sometimes touches 48 °C in summer. It is bordered by almost all regions of Pakistan; Azad and Jammu Kashmir to the North-East, the Indian states of Punjab, KPK to the North-West, Sindh province to the South, Baluchistan and Islamabad (Federal Capital Territory) to the North. **Lahore** is the provincial capital and is located in the East-Central region of the Province, bordered with India. India being the neighbor has a long standing cultural and linguistic links with states of Punjab and

Rajasthan. Traditionally, Lahore has been the capital of Punjab for a thousand of years. It is the hub of administrative, historical, cultural, and economic activities.

Administratively Punjab province has been divided into 9 divisions and 36 districts as explicated in **Figure 2.4**. The divisions are the larger administrative units while districts are subdivision of each corresponding division. Geographically, the whole province can by divide into three main regions (a) Central Punjab; (Lahore, Faisalabad, Gujranwala and Sahiwal); (b) Upper Punjab (Rawalpindi and Sargodha); (c) Southern Punjab (Bahawalpur, Multan and DG Khan). The topography of Punjab is primarily comprises of fertile areas, the waterway valleys and deserts. This includes level plain in the central regions of the province, with the Pothwar plateau lying to the north, and the Cholistan desert to the south.

Lahore is the most urbanized district in the province with 82% urbanized population and highest population density. This urbanization is on the verge of increase and other densely populated districts which lie in north or central Punjab also have the high rate of urbanization (i.e. Rawalpindi, Multan, Faisalabad, Sialkot, and Gujranwala. These districts are all hubs of industrial production either light/electrical appliances or heavy machinery. Faisalabad is famous for Textile products whereas district Jhang has the largest Wheat & Sugar cane production.

The province has an extensive agrarian and industrial based economy which is rapidly growing as compared to other provinces. Punjab Economic Report (2007) reveals that it contributes about 58% to Pakistan's GDP. According to the Government of the Punjab statistics, about 75% contribute to annual food grain production in the country. It is also the most developed province of Pakistan; comprising mainly of like manufacturing industries. Punjab is also major labor contributor because of its most impressive pool of professionals and profoundly skilled (Technically advanced) labor in Pakistan. The province is also enriched with mineral extensive mineral deposits of Gas, Coal, Rock salt and Petrol. The world's second largest salt mine at Khewra, in the district of Chakwal is also in this province.

Useful statistics of some socio-demographic indicators in Punjab are also described here which would be helpful in understanding social aspects of the disease in the region. These indicators are also the part of Millennium Development Goals (MDGs) in the province.

a) **Literacy:** Literacy is a key measure for sustainable development (Iqbal *et al.*, 2003). According to the statistics from MICS survey, Punjab 2011, about 60% of the province population aged 10+ years is literate. This rate is higher than the national average literacy rate by points. The percentage varies with gender; males (68%) and

females (51%). A significantly different literacy rate (74%) has been observed among the population ages between 15-24 years (males 78%; females 70%) (MICS, 2011). Across the province, the literacy rate is not uniform and differs significantly among the districts, gender and urban-rural divides. The highest overall literacy rate was reported from the districts Chakwal (90%) and Rawalpindi (87%), whereas, lowest literacy rate was recorded from the Rajanpur, a district of Southern Punjab.

b) **Unemployment:** About 3% individuals belonging to age group 15+ are unemployed. The rate is almost same in urban-rural areas.

c) **Household Income:** The results from "Household Income and Expenditure Survey 2010-11" indicate significantly different monthly household income among the urban-rural households (HIES, 2011). For example, on the average monthly household income in urban area is Rs. 22,859, while in rural area it is Rs.19,778. The data on expenditure reflects the household characteristics. The results from population Census of 1998 indicate that in rural areas, 53% of the household income is being spent only on food, as compared to 41% for urban households. It is pertinent to mentioned that women in KPK, Sindh and Balochistan province are not well represented in mainstream employment and even access to nutrition, health and education.

d) **Average Household size:** mean household size in the province is 6.3 which are closer to the national average (6.9). Overall dependency ratio in the province was 87.1 with great variation across districts. This ratio is significantly higher in Southern Punjab districts (MICS, 2011).

e) **No. of persons per room:** mean no. of persons per room is 3.7

Overall, most of the resources are diverted for development to Central and Upper Punjab regions, while, Southern Punjab is comparatively neglected for such activities. Thus, it is noticeable that southern parts of Punjab desperately poor and significantly under-serviced.

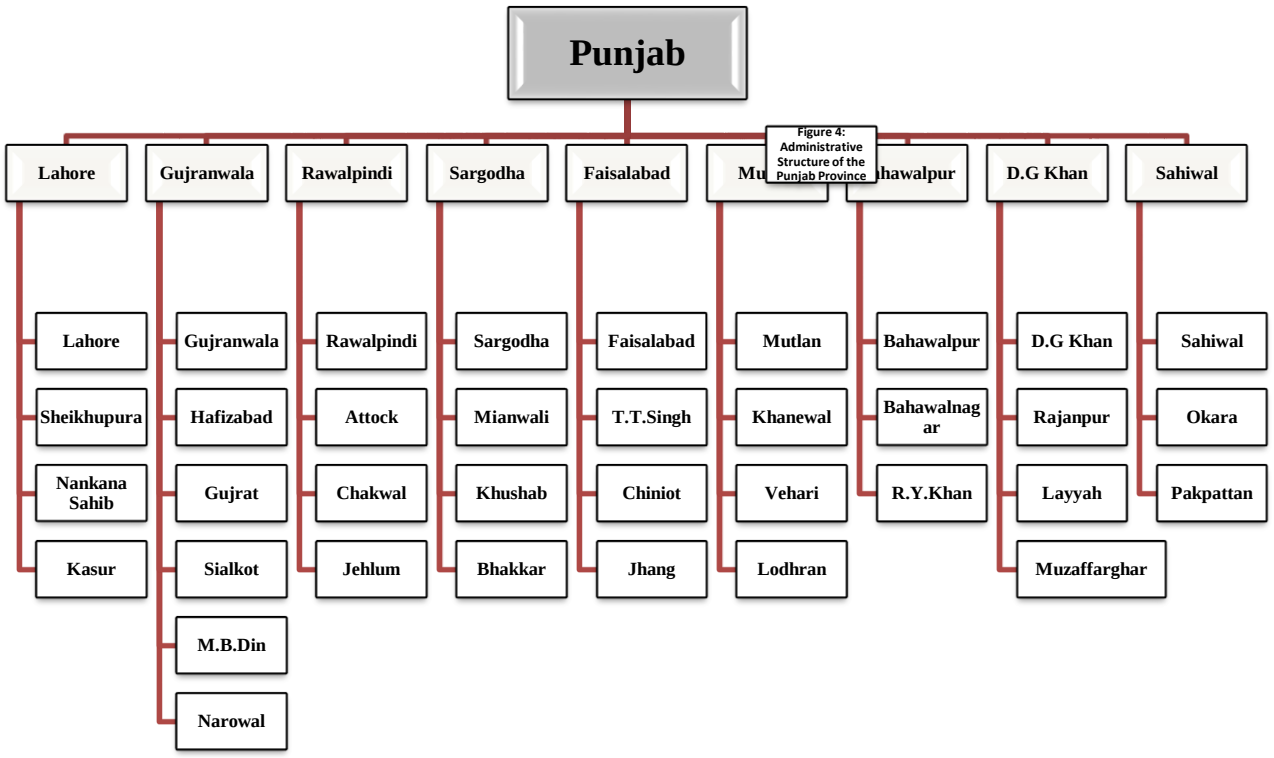


Figure 4: Administrative Structure of the Punjab Province

Figure 2.4: Administrative Structure of the Punjab Province

1.7 Health Facilities

In Pakistan, total numbers of government hospitals are 972 whereas in Punjab this figure is 329 with 39,052 numbers of beds. Moreover, 1,347 dispensaries, 2,540 basic health units and 369 maternity and child health centers are in Punjab. Only 125,034 are the registered medical doctors in the Province out of which 25,633 are specialists in different medical fields. In Pakistan, the population and health facilities ratio manifests that 1,206 persons per doctors and 1,665 persons per hospital bed (Wing, 2011). To compensate the situation private sector is contributing well by accounting for about 80% of patients' burden leaving the remaining 20% at the mercy of public sector. The charges are too high to be borne by the poor patients. Unfortunately, the public sector is over burdened and can hardly fulfill the medical needs of the patients due to meager resources and limited availability of doctors and paramedical staff.

In Pakistan, obviously, the burden of hepatitis is increasing day by day and the federal government has foreseen this endemic situation in the country. Therefore, "***The Prime Minister Program for Prevention and Control of Hepatitis"*** was launched in 2005. The

primary objective of the program was to control morbidity and mortality in the public due to viral hepatitis. Free of charge treatment/diagnosis of HCV was provided to hundreds of thousands people in 150 teaching and District Head Hospitals (DHQs) through-out the country (APP, 2012). But, at the same time, the Health Ministry of Pakistan was devolved towards the provinces in 2011. Therefore, an allocated amount of Rs.684 million could not be utilized in letter and spirit during Financial Year 2011-12 (Ghani, 2012). However, it is believed that this matter is now settled and the program has attained its space of progress and implementation.

1.8 About Sampled Hospitals and Surrounding Population

In this section, a brief introduction of hospitals and the surrounding population which were sampled is discussed.

a) Lahore /Services Hospital

Lahore is the provincial Capital of Punjab and is regarded as Heart of Pakistan. It is 2nd most densely populated city in Pakistan, situated near river Ravi with Pakistan-Indian Wagah border on East. It has an ancient history with rich heritage of the Mughal Empire. Administratively Lahore Division embraces four allied districts i.e Lahore, Sheikhupura, Nankana Sahib and Kasur. The city district Lahore is further divided into nine towns. Many headquarters of international companies lie in this city which helps in generating several economic activities. In overall national economy, Lahore contributes to 13.2% approximately. Thus, it is a hub of political, business, transportation, entertainment and educational activities. Thousands of people are moving to Lahore from different areas of Pakistan for getting better education, job and health services. This city holds many reputable public sector as well as private universities, colleges and hospitals.

Among the public sector hospitals, Services Hospital is a big one declared as teaching hospital for the Services Institute of Medical Sciences (SIMS) on 2002. SIMS is a public sector medical college affiliated with University of Health Sciences, Lahore. It is located on a very famous place at Jail Road, Lahore next to the Race Course Park and Institute of Cardiology. It runs under the administrative control of Government of the Punjab and provides tertiary care health services to the general public. This hospital was initially established on 1958 with 60 beds only, whilst, currently, it has been equipped with 1196 beds with 31 departments, 27 major and 8 minor operation theaters. On the average about 700 out-patients from surroundings and distant areas visit this hospital each day. Efforts are confined

to control major infectious diseases such as Malaria, Tuberculosis, HIV/AIDS, viral Hepatitis B & C, and Dengue with the help of department of infectious diseases. Different major departments such as Endocrinology, medicine, orthopedics, Surgery, Obstetrics & Gynecology and pediatrics are well established. Air conditions are installed in almost every Ward and OPD centers. Importantly, laboratory tests, radio-graphical investigations and examinations are carried out free of cost for all the patients admitted in the hospital and in emergency ward. There are different achievements of this hospital particularly, up-gradation of pathology department with ISO 9002 and ISO 15142 certifications, efficiently development of Hospital Information Management System (HIMS) and Establishment of Endocrine & Diabetic Center. A separate "Hepatitis Clinic" has also been established in this hospital for ready diagnosis and free treatment of hepatitis patients, particularly for Hepatitis B & C patients.

During data collection it was observed that **Hepatitis Clinic** was well managed and record of patients was being computerized properly. Removal of waste material was up to the satisfactory level. However, patients waiting area, toilet, drinking water facilities were seem to be insufficient. Patients were welcomed 6 days in OPD, however, patients who were receiving treatment of Hepatitis C i.e. Interferon Injections had to attend the clinic twice a week only.

b) Gujranwala /DHQ Gujranwala

Gujranwala Division is comprised of 6 districts and is an industrial city situated north-east of Punjab province. It has its own historical and cultural identity. Gujranwala has also many tourist places. It is a neighboring city of Lahore and located only 65 Km away from Lahore. In addition, Gujranwala is extensive contributor in manufacturing and agriculture markets in Pakistan. Literacy rate (10+ years) in women of Gujranwala district is 72% (MICS, 2011). It's vital role is strengthening the economy of Pakistan each day by producing and exporting different things e.g. sugarcane, melons and wheat for international export. It is a major contributor of numerous others item include; textile mills, cutlery industry, electrical equipments, leather industry, machinery, motorcycles and electric fans industry etc. Gujranwala, Sialkot and Gujrat are referred as "golden triangle" of the division due to their active role in national economy. Almost 60% of Pakistan's small and medium enterprises are found in the region. Some of the finest wrestlers of the subcontinent have been originated in this city. Punjabi is the frequently spoken language, however, Urdu and English are also

spoken in offices and educational institutions. About 32 private and public sector colleges, universities are established in this city.

Though different hospitals are providing healthcare services to the people of six districts of Gujranwala, yet DHQ Gujranwala is a single public sector hospital in Gujranwala which is the only hope for the poor people in the region. It is a 450 bedded teaching hospital and equipped with almost all specialties. About 6 Lac out-patients from adjacent localities visit this hospital annually for diagnosis and their treatment (Sonia, 2010). Initially it was established in 1925 and now is attached with a newly established Gujranwala Medical College which was inaugurated on 17th January, 2011. This hospital is located at central city and is easily approachable. Free diagnosis and treatment is being offered here by Government of the Punjab to the Hepatitis C patients. There are 18 indoor wards and one OPD block. Its target population is about 4.4 million (Sonia, 2010). During data collection, the researcher observed that Hepatitis C patients were being treated under Medical Unit and there was no separate Hepatitis C clinic like Services hospital, Lahore. Moreover, no computerized data record of patients was maintained. Proper cleanliness and removal of waste material was not being ensured unfortunately. Similarly, the hygienic condition of wards for admitted patients was not up to the required level. However, the cooperation of medical specialists for data collection was really memorable and acknowledged.

c) Rawalpindi /Holy-Family Hospital

Administratively Rawalpindi is working both as a division and district. As a division it is comprised of 4 districts which include Rawalpindi, Attock, Chackwal and Jehlum. This administrative division is lying in the north region of the Punjab which surrounds the city of Rawalpindi. Rawalpindi district is 2nd most urbanized district of Punjab with an estimated population of 4.5 million in 2010 and covers 5,286 km^2 area. Many people have migrated from different parts of the Punjab and even from other provinces of Pakistan for work in the city. Moreover, about 7,335 Afghans National were reported to be living in Rawalpindi (peopledaily). It is situated northernmost part of the Punjab region and is located very close to the Capital city Islamabad. It is a Pothohar region of Pakistan surrounded by high mountains of “Himalayas” extremities and rich valleys of Mountain Rivers from north and western sides. It also contains the chief hill station in the Punjab i.e. “Murree hills”. The main rivers are the Jhelum and Indus. Major Language of the region is Punjabi (90%), however, other languages including Pothohari, Pahari and Majhi etc are also spoken.

Rawalpindi is referred as military city because Pakistan Army Headquarter is situated here. Many famous places and parks are the big source of attraction for the tourist community here. It also provides the base camp for the tourists who want to visit other hill stations like Murree, Nathia Gali, Abbotabad etc. About 939 industrial units are operating in the district including Kohinoor Textile Milles as the largest unit. Many standard colleges, universities have been established under private and Government administrative control. Similarly, regarding health care facilities government hospitals such as Holy-family, Rawalpindi General Hospital and District Headquarter Hospital serving the local and surrounding districts communities. However, a Holy Family hospital is 400 bedded hospital situated in Satellite Town, Rawalpindi, a big teaching hospital established in 1948. Later on, Rawalpindi Medical College was attached with it on 1977. It is located at easily approachable location. It has almost all specialties including departments of Medicine, Gynecology & Obstetrics, Surgery, Pediatrics and Ophthalmology. The hospital is equipped with advanced equipments and most of the wards are air-conditioned. A rapid expansion is being introduced in the hospital with a 400 bedded brand new block. State of the art equipments including MRI have been installed in wards and Operation Theaters. The most interesting thing is that this hospital has established "Hepatitis Center" separately running under the control of Medicine department where patients of Hepatitis B & C are provided free of cost diagnosis and treatment, sponsored by Government of the Punjab. A computerized data of all hepatitis patients is being maintained. Hundreds of thousands patients are visiting this hospitals each day. Proper hygienic condition, waiting area, sitting place for the visitors and patients is observed in the Hepatitis Center; however, in general hospital, improvement is required in health care facilities. Cooperation of doctors, staff during data collection was really memorable.

d) Sargodha /DHQ Hospital Sargodha

Sargodha is administrative division of 4 allied districts. It acts as division and district as well. District Sargodha is located 206 Km north-west of Lahore city. Total estimated population of district Sargodha as on 2012 is 594,000. Majority of the population belongs to rural set up. It is the 5^{th} largest city in Punjab and famous in agricultural trade and livestock. It produces citrus of export quality. A wide range of crops are being produced here and different animal breeding programs are ongoing. It has flat and fertile plains and is amongst one of the planned cities of Pakistan. The River Chenab lies on the Eastern side while

Jhelum flows on the Western and Northern sides of the city. It became significant place as Pakistan's largest Airbase that is PAF Base Mus'haf is situated there.

Different educational and healthcare centers can be found in the city but also still it was experienced there is more need of enhancements in this area. Sargodha University a public sector university offers best university education for the local community inclusive of PhD degrees. Majority of the patients are member of poor families who can no longer afford treatment expenditures in private hospitals. Presently, DHQ hospital, Sargodha is the main contributor regarding health-related facilities in the area, specifically for poor people. This hospital was attached with newly established Sargodha Medical College (SMC) as a teaching hospital in 2006.This hospital has almost capacity of 600 beds and located in the central location of the city. However, a recent up-gradation has been made within the hospital by constructing new building for OPD patients, Operation theaters, ENT, Urology and general medicine wards. No proper sitting place, water and toilets facilities for the visitors in OPD were observed. Moreover, hygienic arrangements were also pitiable. This hospital was also providing free treatment to the hepatitis C patients but no adequate, organized strategy was opted by the administration to facilitate the patients. This hospital looks over burdened and thus need to upgrade immediately.

e) Faisalabad /Allied Hospital

Faisalabad formerly known as **Lyallpur** is comprised of four adjoining districts i.e. Faisalabad, T.T Singh, Chiniot, and Jhang. Faisalabad district lies in the northeast region of Punjab at about 2.30 hours drive via road from Lahore city. The distance has been reduced adequately because of Motorway development. The name "Faisalabad" was given in the honor of King Faisal, Saudi Arabia. It is the 2nd largest district in the Punjab and 3rd largest metropolitan city in Pakistan, in fact known as Manchester of Pakistan. It has emerged as a major industrial and agriculture center of Pakistan by contributing more than 40% share of total exports from Pakistan. Being the hub of industrial units, it contains about 512 large Industrial units. Similarly, regarding agriculture center, it forms about 55% of Pakistan. The city has been connected with other major cities via all means of transport. A diverse culture is found in Faisalabad because many people from all over the Pakistan come to here for earning better livelihood as a result of working in factories. About 80% of cultivated land is irrigated by the Lower Chenab River.

Certainly, there are different schools, universities and colleges of high standard serving the community in Faisalabad. For healthcare purpose, Allied Hospital is considered as the largest hospital in Pakistan with 1150 number of beds. Additionally, it is a teaching hospital and attached with Punjab Medical College, Faisalabad. Punjab Medical College is amongst the best medical colleges of Pakistan, established in 1973-74. Allied hospital occupies central place on Jail Road, Faisalabad. Hospital is fully furnished with the most recent apparatus and machinery and assertive with all specialties. That's why patients from all surroundings approach even from distant areas to get reliable and cheap medical care. It has an emergency department which runs round the clock. Emergency ward possesses two its independent operation theatres and Radiology lab. Free treatment is being offered for all the emergency patients and the indoor patients. Infect it's a big hospital and receives very high volume of indoor as well as outdoor patients. Outdoor of the hospital remains to be available from 8 am to 2pm on week days.

The majority of the hepatitis C patients visit outdoor and treatment is free of charge for them. However, it was realized that sometime it got challenging to provide interferon injections free to everyone because of stingy resources. However, initiatives, co-operation of doctors and other paramedical staff are attributed.

f) Sahiwal/Civil Hospital

Sahiwal is newly constructed division of Punjab. It has 3 allied districts namely Sahiwal, Okara and Pakpattan. The city is situated between the Ravi and Sutlej rivers, approximately 180 Km away from the major city Lahore. It is the center of Multan and Lahore cities. Population wise it is the 22nd and 14th largest city of Pakistan and Punjab respectively. Majority of the population speak Punjabi language and 99% are Muslims. The region has Agro-based economy because of its fertile land. Major crops of the region are cotton, potato, rice and grains. It is also famous in livestock and Buffalo milk. Although different reputable universities have started to open their regional campuses here but still this region required more educational institutions which can provided better educational facilities to the population. Similarly, in health sector a greater attention of the government is required. However, currently, the Civil hospital is the largest public sector and teaching hospital of the region which is the sole hope of the poor people. It is required to expand this hospital immediately by equipping with latest infrastructure and modern equipments. That's why a sufficient number of patients have to refer to Lahore for better treatment.

g) Multan /Nishtar Hospital

Multan division has its four allied districts which are Multan, Khanewal, Vehari and Lodhran. This division is located in Southern Punjab region and population wise, the district Multan is the 5th largest city in Pakistan covered with an area of 3,721 Km^2. River Chenab flows on its Western side. It is one of the oldest cities in the world and has its own cultural and historical identity. It is famous as the City of Saints or Sufis as many shrines of Sufi saints are found here. Common language is "Saraiki" while variety of Punjabi languages is also spoken. Multan has concrete roads which connect it with other big cities like Faisalabad, Lahore, Karachi and Bahawalpur. The region has mixed culture of North and South Punjab but local people possess loving attitude, therefore, no distinction or hate was noticed among the sects. Majority of the population belong to Muslim religion but some Sikh and Hindu families also reside here. Multan is an important agricultural centre and produces wheat, sugarcane and cotton. Many commercial & industrial centers are also found in Multan. It is famous for its hand-made carpets & ceramics.

The Nishtar Hospital is a tertiary care largest public sector hospital in Pakistan with 1800 beds capacity. It is affiliated with Nishtar Medical College which is an oldest and prestigious medical institution of Pakistan, established in 1951. It is located at a distance of 3 km from the city and the hospital building lies very adjacent to it. The total area including entire campus, hostels and hospital is about 112 acres. There are 29 wards, one OPD, separate Emergency department and one Burn Unit in the hospital. This hospital is serving the largest population of Southern Punjab and even from other provinces of Pakistan.

h) Bahawalpur /Victoria Hospital

Bahawalpur is the 12th biggest district in Pakistan, found in Southern area of the Punjab. It was the home of many Nawabs (rulers) Sutlej River flows around its southern side. It is linked to key cities of Pakistan with concrete roads together with other sources of transportation. The district is enriched with heritage and many historic places. The city carries one of the known natural safari parks in Pakistan. The Cholistan Desert is located on eastern side of the Bahawalpur city which covers about 15,000 km^2 area. Similar to the other districts it also has fertile lands and abundant in agriculture sector. A variety of agriculture goods like cotton, wheat, citrus and rice of international quality are grown in the surrounding region and exported. Other products entail dates and mangoes which have been cultivated

here. It is among the stable regions of the country. Many development projects are ongoing, supported by the Government of Punjab and foreign aid.

The city includes a variety of well-established academic and health-related centers Such as the Quaid-e-Azam Medical College (QAMC) and Islamia University of Bahawalpur. QAMC is the most ancient medical college of Pakistan and was affiliated with Bahawal Victoria Hospital (BVH) in 1970. This hospital is the largest tertiary care facilitation hospital with 1,600 beds, located in Bahawalpur. This hospital was started as a Civil Hospital with an outdoor dispensary in 1876. In 1906 it was given a name of Bahawal Victoria Hospital. Today it is a big hospital delivering healthcare for the more extensive part of people of Southern Punjab using all clinical expertise. Huge numbers of people from adjacent districts and communities visit for their treatment including hepatitis C. Majority of the patients visiting for treatment in this hospital were poor and illiterate. No proper sitting place have been reserved for the visiting patients in OPD, therefore, most of the patients have to wait for their turn outside the floor even in extreme hot and cold days. No computerized data of the patients is being maintained. Cleanliness was also the matter of concern.

i) Dera Ghazi Khan/ DHQ Hospital

Dera Ghazi Khan is famous as D. G. Khan division located in Southern Punjab region. It has four allied districts which are Dera Ghazi Khan, Muzaffargarh, Rajanpur and Layyah districts. It is the center or mid of Pakistan and provides linkage to all provinces of Pakistan. It has unique geographical and cultural identity and enriches with natural beauty including beautiful land, Green Fields, Indus River, Desert and the Suleiman Mountains. DG Khan is also rich in agriculture and produces wheat, cotton, rice, sugarcane and tobacco as the major crops. It is also well known for growing dates. In the past, unfortunately, DG Khan remained neglected area regarding development activities, that's why adequate health and educational facilitates could not be provided to the local inhabitants. Despite the fact, DG Khan has literacy rate of 60% which is relatively greater than the national literacy rate and have produced very eminent doctors and engineers who are serving the nation. Many students have to move to other big cities like, Lahore, Karachi, Faisalabad and Islamabad for higher education. However, now a rapid growth in these areas has been witnessed and sufficient development funds have been diverted to this neglected but important region of Punjab. An international Airport has also been established in Dera Ghazi Khan over 15 Km area.

Regarding health facilities, a civil hospital which is the Divisional Headquarter Hospital (DHQ) is the public sector hospital with 250 beds and serving the local community. Other private hospitals are also trying to provide the best treatment facilities to the population but still needs improvement. Majority of the poor people visit the DHQ hospital which is being expanded up to 500 beds capacity. However, overall picture of public health was pitiable in the region. No proper system has been observed in this hospital through which treatment for hepatitis C patients could make possible on equity basis. No computerized data record of the patients was available.

1.9 Rationale of the Study

In Pakistan, studies reported that lack of knowledge about risk factors of HCV infection is one of major reasons of its rapid spread (Khuwaja *et al.*, 2002, Talpur *et al.*, 2007) and so a thorough understanding of its epidemiology would help to identify modes of disease transmission and develop more effective preventive strategies.

Thus identifying behaviors' which lead to hepatitis C virus infection have become important due to the rapid increase in global prevalence, non-availability of vaccine and expense of treatment (Sharma, 2009, Umar and Bilal, 2012). Although, a variety of the risk factors for HCV are now widely accepted. However, 20% to 40% of infections are acquired in an unknown manner, and we still need to explore (Karmochkine et al., 2006b) and identify other risk factors associated with hepatitis C in different settings or regions of the world. Moreover, many studies describe risk factors for infection in developed countries but comparatively little is described about developing countries such as Pakistan.

To the best of our knowledge, no Punjab-based study has been conducted pertaining to the risk factors of hepatitis C by seeking data from each division in Punjab. Moreover, it is believed that statistical models constitute a very powerful tool for supporting medical data to interpret correctly and then to model and analyze the potential risk factors. For this purpose, the multivariate logistic regression is a conventional as well as widely known technique being employed. However, this technique exhibits few limitations, for instance, the handling of interaction effects and missing values in the data etc. In fact, logistic regression handles interaction in the model very well, but the consideration of higher order interactions is still a problematic subject. Alternative statistical methods such as artificial neural networks (ANN) and classification trees have become popular in recent years by suggesting some alternate solution of these issues. Although these techniques have gained popularity in recent years,

not so much to identify risk factors, but as prediction tools (Xue et al., 1996, Sgourakis et al., 2012, Hong et al., 2011, Heydari et al., 2011, Jajoo et al., 2002). This multicenter, hospital based, case-control study would help to present understanding of these proposed methods and to overcome shortcomings. The data was collected from every divisional headquarter hospital (DHQ) in Punjab and the results from logistic regression were compared / contrasted against other statistical techniques i.e. Artificial Neural Network (ANN) and Classification trees. These techniques would enable us for acquiring much better insight towards the identification of risk factors of hepatitis C disease.

1.10 Objectives of the Study

The main objectives of this study are to:

i. Analyze the data of risk factors of hepatitis C in the (Punjab), a largest Province of Pakistan using descriptive and analytical methods.
ii. Develop logistic regression model (*Gold Standard*) on overall data.
iii. Present a gender wise comparison of risk factors in the region using separate LR models.
iv. A separate comparison of risk factors of Hepatitis C in Urban/Rural settings of Patients using separate LR models.
v. Develop Region-specific LR models for comparing risk factors in Southern and Northern regions of Punjab.
vi. Compare and contrast the results of Logistic regression model with other techniques like *Artificial Neural Networks & Classification Trees* on overall data.

1.11 Hypothesis to be Tested

i. It is hypothesized that risk factors of hepatitis C infection may vary by gender, region and residential location of the patients.

ii. Statistical models are the powerful tool in establishing relationships between disease and exposures. For this purpose, Logistic regression is most widely accepted model in epidemiological studies. However, it is hypothesized that other techniques such as Artificial Neural Networks (ANN) and Classification Tree models which are relatively new can be the useful source in modeling and analyzing pertinent risk factors.

1.12 Importance of the Study

As the global prevalence of hepatitis C infection is rapidly increasing, it becomes imperative to identify the ways/causes of virus transmission for purposing preventive measures. Usefulness of this study is bi-faceted; for medical practitioners and for statistical research workers. Further, this study will also offer an extensive comparison of risk factors of hepatitis C prevailing in different settings and regions of Punjab. Moreover, logistic regression is considered the most common statistical tool to identify risk factors; however, there are alternatives now available. For example, Artificial Neural Networks and Classification Trees) which have gained popularity in recent years. Importance of medical studies is increasing and the researchers should be familiar with various other types of statistical modeling techniques, so that they can apply and select the most suitable method for their studies. Therefore, this research may provide the basis for application of new statistical methods on risk factors data and develop more robust insight for the researcher as well as policy makers. In short, if statisticians and medical research workers put their consolidated efforts in this area, hundreds of thousands people might be spared from this risky illness.

1.13 Study Plan

This dissertation comprises upon six Chapters. The Chapter No.1 includes overview about the topic, brief introduction, rationale, introduction of study area, objectives and significance of the study. Chapter No.2 covers a comprehensive literature review. This chapter was divided into different subsection to get better insight of the previous studies regarding risk factors of hepatitis C in Worlds and the Pakistani's perspectives. In addition, literature regarding statistical techniques *i.e* Artificial Neural Networks and Classification Trees is also incorporated in separate sections. Chapter No.3 contains research methodology adopted to reach on desired objectives of the study. This Chapter embraces study design, sample size determination, data collection procedure, statistical or theoretical frame work and data analysis scheme. Chapter No.4 is the most important part of the thesis as it includes analysis, interpretation and discussion of the results. In this Chapter, descriptive as well as analytical analysis of risk factors is explained. Different logistic regression models are run on different strata of the collected data which includes overall data, male/female data and accounting for classification of urban/rural settings of patients. Similarly, two separate

models are built for North and Southern regions of Punjab to explore risk factors of hepatitis C at their best place. Chapter No.5 introduces the application of Artificial Neural Networks and Classification Trees on the overall data and comparison of results made with Logistic Regression (Gold Standard). Finally, Chapter No.6 gives a comprehensive summary of the findings along with brief conclusion. Further this Chapter offers some useful recommendations and pinpoints limitations of the study. List of references and Appendices are placed at the end.

Chapter 2

LITERATURE REVIEW

A literature review is an explicit or a systematic way to reproduce what has been studied in the past or to identify where the gaps, limitations or area required further improvement on a specific topic. In addition, this segment provides a detailed review of the research methodology that has been employed in the literature on a specific topic. This gives us an insight to find an appropriate study design, sample size, main findings and conclusions of related studies. A great effort has been made to find relevant studies of HCV in different regions of the world. Additionally, the study methods and statistical techniques were also discussed in the subsequent sections. Summary of all the given literature is at Appendix 2.

Literature about risk factors of hepatitis C infection is gathered in local and global perspective.

2.1 World's Perspective

After discovery of hepatitis C virus (HCV) in 1989 as the major causative agent of non-A, non-B hepatitis (Choo *et al.*, 1989), efforts were started to identify its routs or causes of infection. Initially, in United States, it was observed that about 0.4% of blood donors and 1.4% of the general population were HCV positive (Alter *et al.*, 1989, Alter *et al.*, 1999). Later on, Alter *et al* (1990) conducted a descriptive type study at United States by including five main exposures of disease. These exposures were parenteral drug use, blood transfusion, health care employment, multiple partners and household contact with an infected person. The authors further added that incidence of non-A, non-B hepatitis was (average, 7.1 cases per 100 000). A slight decrease has been reported in hepatitis C due to history of blood transfusion. While the proportion of history of intravenous drug use (IDUs) was raised from 21% to 42%. Meanwhile, Girardi *et al*., (1990) and van-den-Hoek *et al*., (1990) reported highest prevalence of HCV in IDUs i.e 67.5% and 74% among the Italy and Netherland population. These studies showed that history of sharing of needles was the significant risk factor in univariate analyses. Girardi *et al*., (1990) also identified that there was no significant association between marital relationship and hepatitis C using Chi-square test. Van-Beek *et al*.,(1994) reported a different prevalence (59%) among the IDUs in Sydney which did not

differs in males and females. A small study by Caldwell *et al.*, (1991) reported that use of alcohol could be another suspected risk factors of viral hepatitis. While another study by Maisonneuve *et al* (1991) reported that out of 117 blood donors, 46.2% patients did not showed history of known risk factors. This study also reveals that blood transfusion, IDUs and marital relation with same gender were the important risk factors in France. Similarly, a study among the blood donors (n=125) suggested that parenteral exposure was the most likely source of HCV infection in the Netherland (van der Poel *et al.*, 1991). The same routs of transmission were identified in epidemiological data of pregnant women and HCV prevalence was reported as 2.28% (Bohman *et al.*, 1992). Kolho and Krusius (1992) also conducted a different study but with similar objectives among the blood donors residing in low-prevalence country, Finland. The authors included demographic, marital relationship related risk behavior and parenteral risk factors in the study and found that patients who had low level of education (OR= 0.3, $p < 0.001$) and parenteral source of infection (OR=7.6, p=0.0000) were at high risk. Other factors such as coupling, age, gender and occupation were found insignificant in univariate analysis. Pohjanpelto (1992) conducted a descriptive study on 160 HCV positive patients in the same country i.e. Finland and discussed only percentage wise status of different risk factors. They found that patients having exposed to blood (89%), 0.6% had a profession exposure with blood, 8.1% of the patients had lived or travelled in Southern countries; particularly in the Eastern Mediterranean and African regions.

An important information was added by the Fuiano *et al* (1992) in a similar study which reported that the route of transmission in more than 50% of the patients with HCV infection was still unknown. This asks for more epidemiological studies to explore hidden/unknown ways of transmission at the earliest. Keeping in view, Brugnano *et al*.,(1992) targeted the different population for evaluating risk of HCV disease in dialysis patients in central Italy. Total 269 dialysis patients were studied and comparatively high HCV prevalence rate (13.3%) was observed in hemodialysis patients as compared to general population. This finding suggested that patients on long term hemodialysis may get the HCV infection. This source of transmission of HCV infection was also conferred by other authors in their different studies (Dentico *et al.*, 1992, Schneeberger *et al.*, 1993).

A case control study to meet the desired objectives was carried out by Kaldor *et al* (1992) from the blood donors' population in Sydney. The cases (n=220) were the confirmed HCV positive patients while controls were negative (n=210). The case to control ratio was almost 1:1. The main finding suggested that a history of injecting drug use (IDUs) was again found

as key risk factor of HCV infection. This study also studied some other risk factors such as tattooing and multiple partners but they could not establish their significance with the disease. Darwish *et al* (1993) performed a study to evaluate risk factors associated with HCV infection in Egyptian blood donors' population. The total n=160 volunteer blood donors were selected out of which 36 (22%) were HCV positive. This prevalence rate was 5-35 times higher in blood donors' population than other countries. Likewise, several demographic factors and medical risk factors were studied by adjusting and without adjusting of some demographic confounding factors. The results depicted that older age, history of injections and history of schistosomiasis (OR = 8.9, 95% CI = 2.35-33.52) were strongly associated with HCV infection. In this study, the risk factor "Schistosomiasis" was first time reported. On the other hand one more risk factor i.e patients with older age was statistically significant while this was actually insignificant in another previous study reported by Kolho and Krusius (1992). Additionally, this would be a first study where data pertaining to risk factors of hepatitis C was analyzed using multivariate Logistic regression. It was apparent from the earlier studies that only descriptive or univariate analysis was performed for such type of studies. One more case-control study carried out in France to examine important risk factors of hepatitis C from the blood donors' population and applied univariate analysis with Fisher's exact and McNemar's tests. All variables having p-value <0.05 were considered as significant. Moreover, the odds ratios of each risk factor along with their 95% CIs were computed. The main significance of the study was the selection and scrutiny of the cases and controls by ELISA method and then confirmatory second-generation RIBA test. As indicated by Kaldor *et al* (1992) cases to control ratio was considered as 1:1. The authors further suggested that despite blood transfusion and injecting drug users (IDUs) the frequency of history of jaundice and elevated ALT were significantly higher in cases than controls.

In the same context, Neal *et al* (1994) presented a case-control study of blood donors in Trent Region, UK wherein cases and controls were interviewed by a structured questionnaire having useful information of risk factors related to personal, occupational, past medical history, travelling histories and family history. There were 150 controls with 2 cases for each control. univariate and unconditional Logistic regression analysis was carried out with odds ratios and their 95% CIs. In this study some new risk factors were introduced, such as: tattooing, ear piercing, travelling, acupuncture and health care workers, in addition of blood transfusion and injecting drug users (consistent with other prior studies (Kaldor et al., 1992, Girardi et al., 1990)). Moreover, ear piercing, marital relationship and travelling

remained insignificant after adjusting and excluding all individuals who have history of injection drug use. A relevant case control study on transfused population was carried out on 138 seropositive by Montagnac *et al* (1994) in Brazil. Eligible donors were tested for Anti-HCV by ELISA method (Darwish *et al.*, 1993) and statistical analysis was performed both by univariate and multivariate techniques (Neal et al., 1994, Kolho and Krusius, 1992, Petrosillo et al., 1995). The results give a new insight into some demographic factors rather than formal medical risk factors. The results suggested that male, older age (Darwish *et al.*, 1993) and nonwhite ethnicity were the potential risk factors in Brazilian blood donors' population.

Another hospital based case control study was performed by Mele *et al* (1994) from Italy for the purpose to establish relationships of certain risk factors with HCV infection. There were 342 cases and 1095 controls. Case to control ratio was observed as 1:3.2 which means that controls are three times greater than cases. Among the cases about 73% patients were jaundiced. The spatiality of this was that the authors first time targeted on general population regardless of blood donors, hemodialysis and Intravenous drug use (Schneeberger et al., 1993, Serfaty et al., 1993, Kaldor et al., 1992). The authors added that blood transfusion and intravenous drug use account for 61% of the infection. Importantly, the analysis was performed by applying logistic regression technique and results suggested some new potential risk factors. The results suggested that surgical interventions, dental surgery, hospitalization, other percutaneous exposures and multiple partners were the potential risk factors. Statistical analysis also showed that by controlling for blood transfusion and drug abuse, surgical intervention and dental surgery were also the additional pertinent factors associated with HCV infection. Contrary to this, a similar study from Spanish population showed that the use of Intravenous drugs, coupling, tattoos, inter-familiar or vertical transmission were not considered as significant risk factor (Garassini *et al.*, 1995). Vertical transmission also includes; transformation of infection from mother to baby. Although, the risk is very low (0.98%) yet this area was under discussion by some authors (Ohto *et al.*, 1994, Moriya *et al.*, 1995).

A first comprehensive, hospital-based case-control study was presented by Chiaramonte *et al* (1996) in Italy to establish association with HCV infection in general population. It was quite apparent from the prior literature that most of the studies were only targeted and employed on blood donors, dialysis patients, pregnant women, and intravenous drug users to explore the risk factors of the disease. Moreover, little was discussed with reference to demographic and socio-economic factors in previously stated studies. In this

study, 500 cases (HCV positive) and 500 age-gender matched controls (Anti-HCV negative/HBsAg negative were scrutinized from two local hospitals by consecutive sampling. This indicates that an equal proportion of cases to controls was taken into the account (Darwish *et al.*, 1993). Importantly, cases and controls having almost similar socio-demographic status were recruited. All data was gathered by an administered questionnaire and analyzed by Conditional Logistic Regression. Results indicated that blood transfusion and drug addiction were the most common and strongly associated risk factors. While additional risk factors such as use of non-disposable needles, hospitalization, and pervious tuberculosis were independent risk factor of HCV. On comparison, history of hospitalization was reported to be insignificant in a previous study by Mele *et al*.,(1994). Moreover, transmission through coupling was not associated with HCV infection at the 5% level of significance and the same finding was supported by Garassini *et al*.,(1995). However, contrary findings were reported by Luksamijarulkul and Deangbubpha (1997) in call girls, Thailand, wherein results showed that women having marital relationship for 9 years or more the risk of transmission of disease increases by 3.5 times. The high Prevalence of HCV antibody HCV positives was reported 9.5% in female prostitutes in Thailand. A parallel study was conducted by Kim *et al* (1996) in Korean population to identify the risk factors of HCV infections. The total 64 cases and 128 controls matched for age (± 3 years) and gender with 1:2 ratio of cases to controls were enrolled. The most strongly associated risk factors were history of acute hepatitis (OR=3.9; 95% CI: 1.4-10.7) and transfusion (OR=2.4; 95% CI: 1.1-5.2). Logistic regression analysis showed that acupuncture, operation, tattooing, tooth extraction, Endoscopy, ear piercing, family history of hepatitis and needle sharing were not associated with infection in the Korean Population. While comparing with a prior study from Italy, surgical operations and dental surgery were the significant factors associated with infection in their population (Mele *et al.*, 1994). However, this study could able to commence some new suspected risk factors *i.e* endoscopy and family history of hepatitis which highlights its importance but at the same time endow with certain limitations of small sample size.

A large sample study was done by Mohamed *et al* (1996) with the objective to identify risk factors of hepatitis C among the Egyptian population. This was a case-control study based on 5071 subjects' cross-sectionally collected from the Ministry of Health Laboratories during certification of free from hepatitis (B and C) to work aboard. All the relevant information regarding the most important known and suspected risk factors was

collected on a pre-designed questionnaire by the physicians. A highest prevalence (31.5%) of HCV seropositivity was observed in overall population and with a significantly lower figure in females (13.2%). The main significance of the study was the application of multivariate logistic regression analysis including effects of both genders and patients area of residence (urban/rural). This study also identified that married couples are at higher risk (OR=2.3; 95% CI 2.1-2.7). Moreover, the individuals living in rural areas also had increased risk (OR=1.7, 95% CI 1.5-2.0). The results further explained that age>25, male, being married, rural residence, injections were significant in the final overall model. In addition blood transfusion was found to be significant in urban females. In Germany, (an developed country) a consecutive sample of 160 HCV positive patients was drawn from the outpatients to investigate possible risk factors of Hepatitis C and found that history of blood transfusion was also most frequent (34.4%) and no source of HCV acquisition was identified in 38.7% of patients. The authors suggested that major factors in these sporadic HCV infections still need to explore (Zeuzem *et al.*, 1996). This was not a case control study but described with frequency and percentage of different risk factors in Anti-HCV positive patients only. A similar study in USA was carried on 301 consecutive patients with chronic HCV infection which illustrates that about 40%-50% subjects have no known risk factor for HCV acquisition (Flamm *et al.*, 1998).

In this regards, Comandini *et al* (1998) completed a case control study for determining the sporadic risk factors of HCV infection in a selected sample of hospital in general population in Rome, Italy. Total 730 subjects including 365 case subjects (positive for anti-HCV on ELISA), and who had denied IDU. These cases were evaluated with an equal number of hospital based controls (Negative for anti-HCV on ELISA). Final multivariate analysis was performed through logistic regression model and results suggested that intravenous injections, minor surgical procedures, and blood transfusion were the common risk factors (Darwish *et al.*, 1993, Mohamed *et al.*, 1996). In addition, delivery in hospital (female only) and diabetes were newly identified risk factors while blood donation remained insignificant.

Balasekaran *et al* (1999) performed an age-gender and ethnicity matched case control study to investigate some new risk factors and possible mode of transmission for sporadic HCV infection in Southwestern United State. The data was collected from the Gastroenterology outpatients' clinics and the cases/controls have no prior history of injection drug use or blood transfusion. Among the total 477 seropositive patients only 58 (12%) meet

the criteria for cases and the same number of controls were recruited with equal proportion of cases to controls (Kaldor *et al.*, 1992, Darwish *et al.*, 1993). All patients were subjected to structured interview to gather relevant information. The data was analyzed by univariate and multivariate technique (Conditional Logistic Regression) and results described that "Heavy Alcohol Intake" (Caldwell *et al.*, 1991) was strongly associated with HCV infection both in the univariate and multivariate analysis. In addition, multivariate analysis showed that history of transmitted disease (OR= 29.3, 95% CI: 1.3–645.6) and tattooing (OR=5.9, 95% CI: 1.8–320.7) were also most strongly associated with HCV infection (Kaldor *et al.*, 1992). However, some studies from other countries, for example, Korea and Spanish, these factors were found insignificant (Kim et al., 1996, Garassini et al., 1995).

Delage *et al* (1999) conducted another matched case control study for evaluating risk factors of HCV infection in Canadian blood donors and found contradictory results with some previous studies. A total 267 confirmed Anti-HCV positive were interviewed telephonically with 1068 controls making case to control ratio 1:4 which differentiate it from the previous studies (Kim et al., 1996, Chiaramonte et al., 1996, Mele et al., 1994). A univariate and multivariate Conditional Logistic Regression was applied on data. By univariate analysis, 23 covariates were associated with anti-HCV positivity while in the final multivariate analysis, only five factors found significant with HCV infection: history of imprison (OR=56.1; 95%CI:11.4–275.7), intravenous drug use (OR=127.5; 95% CI, 26.0–625.0), previous blood transfusion (OR=10.5; 95% CI:4.7–23.2), and tattooing (5.7; 2.5–13) and marital relation (6.9; 3.1–15.2) (Halim and Ajayi, 2000). This study concluded that marital relation might be a contributing factor in spreading HCV infection (Balasekaran *et al.*, 1999). However, the risk factor "having lived in prison" was first time identified and pointed out in this study while other risks factors such as IDU, tattooing, blood transfusion were already well reported (Balasekaran et al., 1999, van Beek et al., 1994, Maggi et al., 1999). These results were further validated in another case control from United States (Murphy et al., 2000b).

This first population based case control study was conducted by Merle *et al* (1999) in France to determine risk factors of contamination by HCV infection in general population. Most of the previous case control studies were hospital based while in this study authors have chosen all cases and controls from the community population living in France. The cases (HCV positive) and controls (2 for each case) were age-gender matched subjects with negative HCV serology living in the same community. Information regarding demographic,

professional, medical and environmental data was gathered through structured questionnaire. Statistical analysis including univariate (Chi-Square or Fisher's Exact test) and multiple logistic regression was employed. The final multivariate analysis manifest that hepatitis C in relatives (OR=4.58) and medical procedure after an accident (OR=1.51) were the newly identified potential risk factors. While others were commonly known risk factors such as: history of transfusion, therapeutic injections, and surgical procedure (Chang *et al.*, 2001). This study concludes that nosocomial and intrafamilial transmission is possible. When these results were compared with another community-based study which was carried out in seven cities in Taiwan (Sun *et al.*, 1999) with a total 272 cases and 282 controls, almost concordant risk factors were captured by multivariate logistic regression, except two, those were newly identified and described in the same paragraph. However, this study exhibits that prevalence of HCV in spouses of index cases (24%) was significantly higher than the general population (4%). A study by Shin et al., (2000) indicated some contrary results that blood transfusion was remained insignificant among the Korean population while other risk factors including surgical procedures and acupuncture were also significant similar to the other studies.

Briggs *et al* (2001) designed a case control study to observe seroprevalence and risk factors for HCV infection in urban veterans. Anti-HCV testing was performed in total 1,032 male patients including 185 (Cases), 847 (Controls) thus making case to control ratio as 1:4.6. A written informed consent was administered before collecting the thorough information on a questionnaire regarding socio-demographic characteristics and potential risk factors. Crude and adjusted odds ratio were computed along with their 95% CI by univariate (Chi-Square test) and Multiple Logistic Regression. The final multivariate Logistic Regression model revealed that IDUs, Tattooing (OR=2.93; 95% CI,1.70-5.08), Blood transfusion as reported by Chang *et al*.,(2001) were the common and routine risk factors whilst some new risk factors included combat job as a medical worker (OR,2.68; 95% CI,1.25-5.60), history of incarceration and relations with a prostitute. Anyhow, history of incarceration was validated in another study conducted from Brazilian Prisoners (Guimaraes *et al.*, 2001) during December, 1993, to January, 1994. A total of 779 prisoners were interviewed to gather information on socio-demographic variables, coupling, and history illicit drug use. Out of total individuals, 41% prisoners were found HCV seropositive.

A very comprihensive community based study was carried out by Habib *et al* (2001) in the Nile Delta region, Egypt, the most prevalent region for HCV infection. In 1997, one half (3,999) of the households of a village in the Nile Delta, were systematically selected to

investigate potential risk factors that related to HCV infection acquisition. All information, pertaining to socio-demographic status, present/past health conditions, and detailed potential risk factors for hepatitis C was collected by interviewing method on a structured questionnaire (Briggs *et al.*, 2001). There were about 23.8% seropositive females and 24.9% males which showed almost similar prevalence rate among both gender. In this study authors worked on a new direction by comparing risk factors in adults and older age groups of patients in a same study. The multivariate Logistic Regression results revealed that among those over 20 years of age, the subsequent risk factors were independently associated with disease; male gender, marriage (Mohamed *et al.*, 1996), blood transfusion, anti-schistosomiasis injection (Darwish *et al.*, 1993), invasive medical procedure (surgery, endoscopy, catheterization, and/or dialysis) and some newly identified exposures were, history of injections from "informal" health care provider, and cesarean section or abortion in (females only). Contrasting with Western countries, intravenous drug use (IDU) seems uncommon in Egypt, particularly in rural area, and even risk of blood transfusion reduced after opting proper screening tactic for Anti-HCV (Medhat *et al.*, 2002). This tends to identify other new dimensions in risk factors identification in Egypt. To continue this work Medhat *et al* (2002) presented a new study with some more detail inclusion of risk factors and analyzing by applying bivariate (Odds ratio with 95%CI) and multivariate (Multiple logistic regression model) by adjusting for age. The authors develop several regression models for patients older than 30 years of age and less than 30 years. Most of the results are consistent with the study (Habib *et al.*, 2001), however, factors like gender male, being married were significant but found insignificant in another study (Medhat *et al.*, 2002).The results did not disclose any association of other risk factors like tattooing, smoking goza (el-Sadawy *et al.*, 2004), barber shave, ears pierching with the anti-HCV which shows some contradictory results.

Alavian *et al* (2002) carried another a case control study from Iranian blood transfusion organization with 193 HCV-positive donors and 196 (HCV-negative). For identification of true cases and controls, Enzyme-linked immunosorbent assay (ELISA) and recombinant immunoblot assay (RIBA) laboratory tests were performed. Data was analyzed descriptively and analytically (Step-wise forward Logistic Regression Analysis) and final results suggested that Transfusion, extramarital activities (Brandao and Fuchs, 2002), non-intravenous drug, endoscopy, and receiving wounds at war were found to be potential risk factors of being HCV-positive with estimated risk as (Odds ratio: 17, , 42.2, 34.4, 52.8, 4 and

5.2, respectively). No apparent risk factors could be demonstrated in 24.5% of the positive cases. Some concordant factors were reported by another study from Barazil (cases=178; controls=356) with same study design and statistical methods (Brandao and Fuchs, 2002). However, this study discussed only few demographic variables (Age and gender) and simultaneously socio-economic variables were not properly studied. In addition, this was added that un-educated individuals were more likely to be HCV positive and marital relationship with a positive partner demonstrated OR =3.7 (Delage *et al.*, 1999).

Kim *et al* (2002) carried out another case control study from Korean population in different context to identify the risk factors of hepatitis C infection according to the hepatitis C virus genotypes. The Cases (n=178) were positive for HCV-PCR and controls (n=226). All information related to epidemiology of HCV infections were gathered for all participants and data was analyzed by univariate and multivariate analysis. Factors that were significant ($p<0.05$) in univariate analysis, further included in multiple logistic regression model to establish any observed association with hepatitis C infection. Final multivariate analysis results suggested that endoscopy (OR=2.80) and blood transfusion (OR=2.90) were found to be risk factors for HCV genotype 1b against the community controls. While, interestingly, the same risk factors were found significant for genotype 2a which depicts that the risk factors associated with HCV infection are identical among different genotypes. Moreover, these risk factors (blood transfusion & endoscopy) were already reported in another previous case-control study from Korean population by the same author (Kim *et al.*, 1996)

A study from the neighboring country, China was carried out from elder population of Taiwan by Lin *et al* (2003) to find prevalence of HCV and its associated risk factors. Total 1316 individuals including 607 male, 790 female having (age: 62 ±12 years old) were recruited between October 1999 to June 2000. The prevalence of HCV seropositivity was observed as 5.1%. Multivariate logistic regression analysis suggested that age>60, surgical history (Mele *et al.*, 1994) and use of non-disposable syringes (Chiaramonte *et al.*, 1996) were the common risk factors in Taiwan province.

A cross sectional, seroepidemiological study was presented by Mishra *et al* (2003) to gather potential risk factors of hepatitis C infection from the male and female veterans of the North Florida and South Georgia who were visiting veterans' affairs hospital. The total 274 study patients were recruited from the outpatients' clinics during 12-month period from April 1999 to April 2000. All data was collected on a self-administered questionnaire and analyzed by univariate analysis; Chi-Square test with Yates' correction, for binary independent

variables and Student's test, Mann-Whitney U test continuous variable. The dependent variable was binary in nature (Yes/No) therefore; multivariate analysis (Logistic regression with Odds ratio and their 95% CI) was also applied using SPSS software. The final results indicated that following risk factors: use of illicit drugs (OR=3.7, 95% CI: 1.3–11.8) (Guimaraes *et al.*, 2001), incarceration (Briggs *et al.*, 2001) and a Low income (<US $10,000) were the most strongly associated with hepatitis C infection and authors suggested that these factors should be addressed properly to develop preventive strategies in the region. On the other hand, certain following risk factors were expected to be significant but remain insignificant in this region e-g (marital status, level of education, blood transfusion, body piercing, tattooing, contact with a female prostitute/ STD and acupuncture. Some factors such as body piercing and tattooing were also found insignificant in another study from Egypt (Medhat *et al.*, 2002) which support to these findings. However, these risk factors were found significant in various other studies which were conducted from different parts of the world (Mohamed et al., 1996, Habib et al., 2001, Kolho and Krusius, 1992, Delage et al., 1999, Neal et al., 1994, Briggs et al., 2001). Similarly a study from urban population of Haiti showed that injecting drug users (IDU) and multiple partners were the ongoing risk factors (Hepburn and Lawitz, 2004) and almost matching risk factors were identified by (multivariate analysis) among blood donors, northern Thailand in a prospective case-control study of 166 matched sets (Thaikruea *et al.*, 2004). A matching case control study from Hawaii (Lasher *et al.*, 2005) showed that IV drug use, tattoos, incarceration, blood transfusion, acupuncture, dental surgery were the potential risk factors in Hawaii region. In addition HIV infection was also found associated with HCV infection. A similar case control study was conducted by Hand and Vasquez (2005) to evaluate potential risk factors of HCV infection on the Texas–Mexico border. The cases (Anti-HCV positive) and controls (Anti-HCV negative) were enrolled during 2 years period. A thorough telephonic survey was conducted from cases (n=320) and controls (n=307) on a structured questionnaire and the multivariate analysis showed the only injection drug use (IDUs), tattooing (Pérez *et al.*, 2005), and blood transfusion were established risk factors for HCV infection. This illustrates that risk factors pertaining to hepatitis C infection vary by region and population, which demands for appropriate community interventions focusing those with well-established risk factors as a preventive measures.

A study by Zaller *et al* (2004) was presented among blood donors in Georgia, USA to evaluate risk factors of hepatitis C infection. Total 553 volunteer blood donors were recruited from three Georgian cities between October, 1997 and June, 1999. Among these only 43 (7.8%) subjects were found anti-HCV positive and 509 anti-HCV negative. Thus case to control ratio was 1:12, which is running with surprisingly large number of controls compared to cases. The descriptive and univariate analysis was performed before applying multivariate logistic regression analysis. On running multiple logistic regression model, it was noticed that standard errors of regression coefficients of many variables were pretty large because of this inappropriate case to control ratio. Anyhow, to avoid the situation, logistic regression models were repeatedly run by eliminating certain confounding variables and final model revealed that among the demographic and medical risk factors; blood transfusion (OR= 25.9, 95% CI: 3.2–210.9); surgical procedure (OR= 148.4, 95% CI: 26.9–817.4) history of hepatitis (OR= 25.9, 95% CI: 4.6–145.5) were the strongly associated risk factors with hepatitis C infection. Patients' Education was appeared statistically insignificant in the model but still assumed as considerable important variable.

Mendes-Correa *et al* (2005) also presented another case control study to find risk factors of hepatitis C among the HIV patients in Brazil. Before going into the thorough survey, the sample size was estimated using Epi-info software and considering equal proportion of case and controls, 95% confidence level, and 80% power of the test. To our best knowledge, this would be a first study in the context of hepatitis C risk factors, wherein sample size was properly estimated. Data based on socio-demographic, patients' history, marital relation and history of matting was analyzed by both univariate (Chi-Square test) and multivariate (Un-Conditional Logistic Regression). All variables that have shown (p-value<0.20) were further included in multivariate model to examine any independent association of risk factors with hepatitis C infection. The final multivariate model exhibited that older age>50, blood transfusion, IDUs, illicit drug user, and marital relationship with a history of injecting drug user were the common and ongoing risk factors. A different study group was targeted by Yazdanpanah *et al* (2005) and designed a case control study to explore risk factors of hepatitis C transmission among health-care professionals in five European countries. Data from 60 case patients and 204 controls, related to HCV epidemiology was gathered on standardized questionnaire which was translated into relevant language and interviewed by the national investigator. For statistical analysis, crude odds ratio were computed by univariate logistic regression and all those variables which have shown (p-

value<0.25) were submitted for final multivariate logistic regression model. Backward step wise regression method was performed to scrutinize the variables at 5% level of significance. Moreover, the goodness of fit test was also applied in this study, which enhances the capability of this study compared to others. In this study, it was observed that occupational transmission through needle stick and deep injury augment the risk of hepatitis C infection in health-care workers. The conclusion of this study was also supported by a case series study in Poland (Chlabicz *et al.*, 2006) wherein consecutive sample of patients pertains that occupational hazard found in 9% healthcare workers and a careful history exploration can help in recognizing known risk factors in 59% of patients. In a study group at least one risk factor was found in 98.4% of the patients (Karaca *et al.*, 2006).

A case control study from our neighboring country, Iran to investigate possible risk factors for the HCV acquisition in Khuzestan Province, South-West of Iran was carried out with 254 cases and 260 controls (Hajiani et al., 2006a). The cases were selected from HCV positive individuals referred to the Hepatitis Clinic. In Iran, the latest studies showed that HCV prevalence rate lies between 0.12-0.89 percent in general population. The results from univariate analysis showed that blood transfusion, intravenous drug users and hemodialysis were the potential risk factors in Iran. This study was lacking suitable statistical analysis because no multivariate analysis was performed. Anyway, risk factors found in this study were consistent with other studies. Similarly, a descriptive study in Turkey was carried out from the patients, who were admitted, Istanbul Medial hospital during 1996 and 2002. This study found that blood transfusion (39.7%) and surgery (98%) were the most common risk factors on the basis of percentages (Hajiani et al., 2006b, Kim et al., 1996) . But this study only concludes on the basis of descriptive statistic and never applied any test of significance which expressed inadequacy in results.

Karmochkine *et al* (2006a) conducted a very comprehensive, hospital based, large case control study among the adult population of French people, similarly, “To find the risk factors for hepatitis C infection in patients with unexplained routes of infection “.The total sample size of this study was 1250 including (500 cases and 750 controls) showing case to control ratio *i.e* 1:1.5. For cases, only known HCV seropositive patients with age =>18 years were enrolled and got signed consent. Authors struggle to explore new exposures associated with HCV infection and found that among the patients with un-explained modes of transmission, 73% can be explained by the newly identified risk factors of hepatitis C infection. This study further stated that in 20%-40% of the HCV positive patients, no clear

rout of transmission was identified which demands for new diversions in the on-going studies which can explore further hidden exposures in different aspects. The multivariate analysis incriminate 15 different independent risk factors and found significantly associated with HCV infection, for example, wound care (Alavian *et al.*, 2002), intranasal cocaine use, diathermy, gamma globulin, diathermy, varicose vein sclera-therapy, contact sports, beauty treatment and so forth. This study also confirmed and validated some previously recognized (digestive endoscopy, hospitalization, and acupuncture) (Alavian et al., 2002, Mele et al., 1994).

Another case control study from the smallest country of South Africa, the Republic of Tunisia was carried out by Ben Alaya Bouafif *et al* (2007) with only 57 HCV positive cases (61.63 ± 14.84:mean age) and 285 controls (mean age: 60.95±14.66) were recruited. The Logistic regression analysis showed that only traditional risk factors: history of invasive procedures (AOR=2.53; 95%CI:1.21-5.29) (Habib *et al.*, 2001) and intravenous drug injections (AOR=1.96; 95%CI:1.02-3.8) were significantly associated with HCV infection. These results suggested that nosocomial rout of HCV infection in Tunisia was frequent. Almost similar traditional risk factors were reported in a case control study (Kerzman *et al.*, 2007) from two groups (immigrants and non-immigrants) in Israel with the same study design and statistical analysis. An entire different result was observed in a descriptive study from pregnant women in our neighboring country India, wherein data pertaining to risk factors of hepatitis C among the pregnant women was scanty. The study showed that prevalence of hepatitis C in pregnant women was found to be 1.03 percent. The study also observed that out of total 84 anti-HCV positive women, about 61.9% did not showed any identified risk factor and even no significant association was found between the know risk factors and the disease.

A first community based case control study in rural population of Vietnam was completed by Nguyen *et al* (2007) to find prevalence and associated risk factors of HCV infection. A multistage sample of total (n=837) was collected by acquiring relevant information regarding to demographic and potential risk factors by face to face interviewing method. In statistical analysis, Chi-Square test of association was applied to see any observed association between the individual risk factor and outcome variable. Moreover univariate and multivariate logistic regression was employed for assessing significant risk factors. All variables in which p-value<0.25 was evaluated those only included in final multiple logistic regression model (Yazdanpanah *et al.*, 2005). Thus final results suggested that at 5% level of

significance only hospitalization (Chiaramonte *et al.*, 1996) and having tattoos (Briggs et al., 2001, Delage et al., 1999) were the independent predictors with HCV infection in rural veitnam. Wolff *et al* (2008) conducted a different study for identification of risk factors of hepatitis C among the HIV cohort in Brazil. The cases (n=227) and controls were enrolled in the study and multiple logistic regression analysis was performed. The final results suggested that age between30-49 years, elementary school education, lower income status, sharing personal hygiene, IDU and crack cocaine were the independently associated risk factors with HCV infection among the HIV co-infection.

Hepatitis C is also the public health problem in our neighboring country, China. The prevalence of HCV infection China was reported as 3.2% in a survey, 1992, however, unlike the other parts of the world, the prevalence rate also varies within different geographic regions in China. In this regards, a case control was performed by Liu *et al* (2009) in a Henan province, China. Total 69 community based cases and 207 matched controls were selected for collecting detailed socio-demographic characteristics and potential risk factors related to HCV infection. The multivariate analysis showed that blood transfusion, IDU, injections were the traditional risk factors of hepatitis C infection while in addition to these; a newly identified exposure was esophageal balloon examination. This study concluded that unregulated medical procedures may contribute in substantial risk for HCV spread in Republican of China. Similarly, in a cross sectional survey of our neighboring Muslim country, Iran the common and known risk factors of HCV infection among addicted prisoners were reported by Zakizad (2009). The author suggested in multiple logistic regression that only tattooing, multiple partners and history of surgery were the common risk factors. However, duration of imprisonment (mean 48 months), length of alcohol consumption, route of drug administration, sharing of needles and razors remained insignificant in multivariate analysis. This study concluded that prevalence of HCV infection among the addicted prisoners in Iran was very high because of unsafe behaviors and sharing of contaminated utensils. On the other hand in another study imprisonment was the significant risk factor (Pérez *et al.*, 2005).

Another study by Gheorghe *et al* (2010) was aimed at describing the seroprevalence of HCV infection and its associated risk factors in Romania. A nationwide cross-sectional survey was conducted among the adult population during (2006-2008) in Romania, using multicenter stratified random cluster sampling. The prevalence of HCV infection (3.23%) in Romania was similarly reported as in China (Liu *et al.*, 2009). The multiple logistic analyses

depicted that following important risk factors were found to be insignificant in Romania population; organ transplantation, imprisonment, hemodialysis, tattooing, dental therapy, acupuncture, body piercing, attending beauty salon, history of STD, sharing tooth brush or blades, alcohol abuse and multiple marital relation. At the same time, these aforementioned risk factors were identified as significant factors in other studies from different regions of the world (Karmochkine et al., 2006a, Neal et al., 1994, Caldwell et al., 1991, Hepburn and Lawitz, 2004) which showed regional, settings and population wise variation in the significance of different exposures. On the other hand, in this study significant risk factors were: exposure to blood products, accidents/trauma, occupational hazard, injections at home, and intravenous drug user (Merle et al., 1999, Neal et al., 1994, Briggs et al., 2001). While among the young Thai men only these two (injections and intravenous drug user) were the potential risk factors (Jatapai *et al.*, 2010, Mostafa *et al.*, 2010). These results suggested a way to prevent HCV infection by spreading proper campaign in their respective communities.

Awadalla *et al* (2011) conducted a cross sectional study with (n=1000; HCV positive 168, HCV negative 832) among the Egyptian blood donors. This study endows with comprehensive information on socio-demographic, social, occupational hazard and behaviour possible risk factors HCV infection in the area. The collected data was analyzed descriptively and bivariate analysis using Chi-Square test of association at 5% level of significance and result found that poor socio-economic status, being married, Accidental wound, anti-schistosomal treatment and so forth were the potential risk factors. On the other hand, a cross sectional study showed that the risk factors for HCV acquisition in Nigeria have not been properly described (Obienu *et al.*, 2011). In this study researcher tried to evaluate risk factors associated with HCV infection in Nigeria with sample size (n=360), Anti-HCV positive=17 and Anti-HCV negative=343and found that no any risk factor was found significant by applying Chi-Square test of association at 5% level of significance. These results do not look satisfactory because of poorly stated case and control ratio and inadequate study design and statistical analysis. This demands to conduct some new studies in Nigeria with good and comprehensive study design to evaluate epidemiological reliable data.

A similar case control study from China was also conducted by He *et al* (2011) with 305 cases and 610 controls (Ratio 1:2). The sample was collected from Chengdu Blood Center, a largest specialized blood collection center in western China and after collecting relevant data about socio-demographic, lifestyle, and medical risk factors, it was analyzed by univariate and multivariate analysis. The multivariate logistic regression, forward step wise

variable selection method, analysis showed that, no significant difference was observed between the cases and the controls for the socio-demographic variables. In addition, marital status and occupational distribution were also remained insignificant. In contrast, the independent risk factors of HCV infection were observed as: blood transfusion (Romero-Figueroa *et al.*, 2012), razor sharing (OR=29.16; 95% CI: 12.89–66.00), acupuncture, hospitalization, family history of hepatitis, injections>10 years earlier, and ear piercing and dental treatment. Additional significance of this study was that author also evaluated population attributable risk of each associated risk factor. In this study, simply, Spearman's Rank correlation was determined to find any collinearity among the independent variables. In China the risk of hepatitis C transmission through unsafe injections have been reduced at maximum because of rapid economic growth and strict government public health policy (YIN *et al.*, 2004). In the same time, a meta-analysis was performed on 25 case control and cohort studies conducted in China for exploring primary risk factors regarding HCV infection (Su and Wang, 2011). For this purpose, the pooled cases (n=4370) and controls (n=8606) were accumulated for final analysis. The same univariate and multivariate logistic regression was applied to gather primary risk factors. The most significant factors were found to be blood transfusion, surgeries, intravenous drugs, mating with injecting drug user and STDs. These finding were supported by another meta-analysis, cross sectional studies conducted in two Brazilian Inland regions (Souto *et al.*, 2012). Overall, the study corroborated injecting drug user was the main risk factor for spreading HCV infection in the Brazilian region. A retrospective study in Spain further explained that breast feeding does not pay any role to spread the HCV infection in the newly born babies (Madurga Revilla *et al.*, 2012). Another lasted case control study was published by Kandeel *et al*., (2012) to identify potential risk factors of acute hepatitis C infection in Egypt. The eligible cases and controls were enrolled from the two major hospitals in Egypt Between June 2007 and September 2008. The cases (n=86) and controls (n=287) were estimated at 80% power of the test, 3% prevalence and case to control ratio (1:3). The final multiple logistic regression analysis reveal that reuse of syringes, imprisonment, IV fluid in hospital, minor surgical operations, hospitalization etc which concluded that, In Egypt, health-care settings have an ongoing impact on spreading HCV infection. Among the intravenous drug users (IDUs) in Iran, the main risk factors were tattooing, history of imprisonment and sharing of needles (Nokhodian *et al.*, 2012). Another latest study from America reveals that tattooing was the most significant risk factor of HCV infection with OR=3.81 (Carney *et al.*, 2013).

From the above discussion some factors which are common to a group of researchers are older age, H/O un-safe injection, blood transfusion, injected drug use and invasive surgery. Other factors such as dental surgery, H/O hospitalization, tattooing, and family H/O hepatitis are also the traditional factors in most of the aforesaid literature.

2.2 Pakistan's Perspective

In this section, only the relevant studies based on Pakistani's Perspective were discussed to enlighten about the differences and similarities existed in the epidemiology of HCV infection in the region as well as other regions of the world. This would also help to explore shortcomings or the area which needs further attention of the researchers to bring improvement to control the infection in Pakistan.

In this regard, Luby *et al* (1997) conducted a first case control study after the discovery of HCV in 1989 in Hafizabad District of Punjab, Pakistan to determine the prevalence and routs of transmission of HCV infection. A sample of (cases: n=15) and (controls: n=67) was collected out of 504 randomly selected households. The prevalence of HCV was found to be 6.5% in the region. In this study, authors studied different known or community based risk factors of hepatitis C infection which were analyzed using Chi-Square test, Fisher's Exact test, Mann Whitney or t-test (where applicable). The results suggested that among the list of different risk factors e.g. blood transfusion, tattooing, injecting drug users, barber shave (Gheorghe *et al.*, 2010), sharing razor, ear/nose piercing, sharing tooth brush, dental treatment, marital relation, sharing room; only the injections history was the significant factor at 5% level of significance. Remaining other factors were still found insignificant. Although some of these risk factors were identified as significant in some other international studies (Neal et al., 1994, Gheorghe et al., 2010, Liu et al., 2009). However, this study was carried out with small sample size and even statistical analysis was not much sound, therefore, the results were not up to the required level. Likewise, another case control study by Bari *et al* (2001) was conducted from male adults in Rawalpindi-Islamabad, Pakistan with little extension in sample size and with better statistical analysis. The cases (n=57) and controls (n=180) were selected from 9 different hospitals in Rawalpindi/Islamabad region. The univariate and multivariate analysis was applied to examine significance of different risk factors with the hepatitis C infection at 5% level of significance. The final resutls suggested that for adjusting age; therapeutic injections (OR=2.8, 95% CI: 1.1-7.1) (Khan et al., 2008b), daily shave and armpit shave by a barber were the potential risk factors among the male adults in Rawalpindi/Islamabad community.

This study concluded that only the sterlized/disposable syringes should be used and the community people be guided to avoid the use of contaminated instruments in the barber shops. Interestingly, the most common and traditional risk factors such as blood transfusion and injecting drug users were not found signficanct in these studies. Meanwhile, a cross sectional survey by Janjua and Nizamy (2004) was carried out from the barbers in the Rawalpindi/Islamadad locality just to know what sort of knowledge and practices they have about the trasnmission of hepatitis C and B virues. The percentage results suggested that about 13% barbers knew that hepatitis C is a disease of the liver and could be transmitted by sharing razors. The authors further observed majority of the barbers where totally illitrate and had habits of reusing of razor for 46% of clients whilst cleaned with antiseptic solution for only 11.4%.

A similar cross sectional study from Karachi, Pakistan presented differernt findings using logistic regression model that household contact and sharing of tooth bruses were the common risk factors among the thalassaemic HCV seropositive children, Karachi (Akhtar *et al.*, 2002). Another case control study conducted by the same author (Akhtar *et al.*, 2004) to assess the significant risk factors associated with hepatitis C among the volunteer blood donors in Karachi, Pakistan. For this purpose, a consecutive sample was drawn from two blood banks in Karachi between 1998-2002 were enrolled. The cases (n=80) and controls (n=160) between 18-64 years and congregate other eligible criteria were enrolled for through interview. About 88% of controls and 83% of cases were of atmost 35 years of age. The case to control ratio was 1:2 and sample size was estimated at 80% powere of test and 5% level of significance; with the hypothesized risk factors having prevalence of atleast 3% in respective population. The author performed both univariate and multivarite logistic regression analysis in addition to descriptive statistics. Among the socio-demographic and clinical risk factors; only those variables were included in final multivariate logistic model having ($p<0.20$) (Mendes-Correa *et al.*, 2001). The final backword stepwise logistic regression model reveal that hospitalization, theraputic injections and multiple partners were the potential risk factors of hepatitis C among the blood donars living in Karachi. These are the similar factors as reported in similar studies in developing countries (Mohamed *et al.*, 1996, Brandao and Fuchs, 2002, Nguyen *et al.*, 2007). However, the overal epidemilogy of HCV seropositivity among blood donors in Pakistan somewhat different then other developing countries (Akhtar *et al.*, 2004). It was furhter illustrated by Janjua *et al* (2005) in a population based cross sectional survey to estimate annual number of injections per person and their associated cost

in province Sindh of Pakistan. It was noticed that about 76% patients have the history of injections for curative purpose received from the dispensors and 67% from the qualified general practitioner. This study concluded that injections are overused in Pakistan that should be minimized up to the maximum extent to curtail the spread of disease. The authors further explained in their one more study from the Sindh province that individuals who had visited unqualified dispensors/practitioners were more likely to received injections just to get quick relieve from their disease symptoms (Janjua et al., 2006a).

A case control study was published by Shazi and Abbas (2006) to compare and identify the risk factors of hepatitis B and C virus in patients visiting Gestroentrology clinic, Karachi, Pakistan. A convenient sample was drawn out of eligile patients; 63 cases, 44 controls identified by common blood examination method *i.e* ELISA method of HCV seropositivity. On comparing hepatitis C risk factors with the control group, the potential risk factors were found as; low education status, blood transfusions (Neal *et al.*, 1994, Chiaramonte *et al.*, 1996, Chang *et al.*, 2001), occupational exposure (Chlabicz *et al.*, 2006), therapeutic injections (Akhtar et al., 2004, Luby et al., 1997), barber shave (Bari *et al.*, 2001) and intravenous drips. This study emphosis that the proper roper screening of blood before transfusion, use of 1/V drips and barber shaves should be focused. However, this study was carried out with small sample size and poor statistical analysis.

Another case control study was conducted by Ijaz and Akhter (2007) to evaluate risk factors of HCV infection in Lahore; a provincial capital of province Punjab, Pakistan. For this study a convenient sample of 330 patients (including 135 cases and 195 controls) was collected from the major hospitals of Lahore city. The authors applied Chi-Square test, odds ratio along with their 95% CIs for the analysis. However, no multivariate analysis was used to find potential risk factors associated with HCV infection which is lacking of this study. Similarly, authors did not described the selection criteria of cases and controls properly and even inclusion/exclusion criteria was also missing. On the other hand this study included medical history of predisposed factor and found significant in the Chi-Square analysis at 5% level of significance e.g Pruristis history, Malena histor and Encephlopathy etc. The most strongly associated risk factor of hepatitis C infection was injection history which is most repeated risk factor of HCV infection in Pakistan. Among the socio-demographic factor, gender was remained insignicant while patients belonging to rural area were repoted as more likely to HCV infection than urban patients. Additionally, the authors also suggested that the interaction effects of the risk factors should also be identified in future studies. Some other

factors such as surgery, barber shave and blood transfusion were also found significant but still it is noticed that this study have not applied multivariate analysis, therefore, these factor need further investigation with proper statistical analysis to improve and validate the results. Meanwhile, this improvement was done by Abbas *et al* (2008) who carried out a study to determine "Prevalence and mode of spread of hepatitis B and C in rural Sindh, Pakistan". In this study a cross sectional survey of 843 subjects belonging to 174 different families was undertaken by systematic random sampling. Simple logistic regression model was performed first to find any association of HCV infection with individual independent variable; odds ratios (OR) and 95% CI were also computed. All those variable having ($p<0.25$) were selected for final multivariate logistic regression analysis (Yazdanpanah *et al.*, 2005) to explore most important risk factors associated with outcome at p-value < 0.05. The final results reveal that most stongly associated risk factors for hepatitis C were; age>16 years, dental procedures, history of hepatitis, and atleaset 10 injections per year with their estimate risk (odds ratios: 3.7, 2.1, 2.4, 1.8 and 2.9) respectively. In addition, this study also highlighted that intrafamilial and household clustering (parent to child, p=0.001; sibling to sibling, p=0.046) might be possible. These findings are supported by a case control study from France (Merle *et al.*, 1999).

A prospective study by Idrees *et al* (2008) was conducted from the province Punjab, the province under study with the aim to find prevelance of hepatitis C infection among the general population. This would be the first study after Luby (1997) with community based sample and better study design. The overall prevalence of HCV infection was established to be 14.63% while this figure was significantly different in males (15.09%) subjects compared to females (12.30%). Besides their main objective, the authors also evaluated significant risk factor in the region by applying univaritate and multivariate analysis. The results showed that injected drug use (adjusted OR=6.6), needle prick (adjusted OR=2.2), re-use of syringes (adjusted OR=1.7), blood transfusion (adjusted OR=5.9) and age> 35 years (adjusted OR=1.3) (Darwish *et al.*, 1993, Wolff *et al.*, 2008) were important risk factors for HCV infection. These factors were also fond significant in different international studies by different authors and settings or regions which cherished us that some factors like injecting drug users and blood transfusion were common even in developed countires , for example in Italy, Canada and France (Mele *et al.*, 1994, Delage *et al.*, 1999, Merle *et al.*, 1999) and so forth. The author further concluded that, In Pakistan about 70% of the cases were obtained through reuse of syringes and general surgery (Idrees and Riazuddin, 2008)

while on the other side about 20.35% subjects have not any clear mode of transmission in our country (Muzaffar *et al.*, 2008) which point out the alarming situation and insist for further epedimiological studies. In this regards, Ali *et al* (2009a) conducted a meta-analysis regarding prevalence and mode of transmission in Pakistan by accumulating results from different published studies from Pakistan.The authors, extracted 6 different modes of HCV transmission in Pakistan including; Accidental needle stick in healthcare settings, Receipt of blood and blood products, Injection drug users (IDUs), Occupational exposure, barber shaving, and Household contacts/intrafamilial transmission. These sources of HCV transmission are also well reported in international studies, however these risk factor varies by region, setting and target population. Hence to make well established preventive strategies it is blieved that separate studies are required to explore potential risk factors pervailing in different communities or settings of the world.

A hospital bases case control study was conducted by Ghaffar *et al* (2009) to evaluate risk factors of hepatitis C among the women of reproductive age, in Quetta, Balochistan province of Pakistan. The womens of age 18 to 40 years from two leading government hospitals where people belonging to both urban and rural settings reached for getting their treatment. For Cases only Atni-HCV positive patients on ELISA test were considered while controls have a proved evidence of Anti-HCV negative . Total 216 patients were selected with equal case to control ratio and age-wise there was no differece. All data was gathered on a structured questionnaire and analyzed by bivariate (Chi-Square and multiple logistic regression. Descriptive statistics showed that about 51.5% cases were illitrate and belong to poor families. More than 60% controls belong to urban area while 45.6% cases belong to rural communties. The final multivariate results indicate that history of injections, family history of jaundice and previous surgeries were the independent risk factors of HCV infection in womens. Other factors such as caesarean section, history of abortion and D&C remained insignificant. Anyhow, these risk factors were firs time included in any study from Pakistan and among females only, which make out its importance. For male only, another separate hospital based case control study was carried out by Qureshi *et al* (2009) from Karachi, Sindh province of Pakistan. In this study , authors tried to compare the risk facotors of hepatitis C & B assuming that both types of viruses contain almost similar risk factors. For sample size a consecutive sample of (n=1446) patients were enrolled from the two reputable hospitals in Karachi, City by assuming equal proportion of cases to controls (Ghaffar *et al.*, 2009). Similar to other studies, a logistic regression model was applied separately to identify risk

factors for both virues *i.e* Hepatitis C & B. The resutls related to hepatitis C explained that dental surgery, blood transfusion, I/V and I/M history of injections, family history of hepatitis, surgery (Kim *et al.*, 1996, Ghaffar *et al.*, 2009), ficial barber shave and so forth were the main risk factors contributing in spreading hepatitis C virus among the males only in Karachi. These two separate studies on males and females suggested that some risk factors of hepatitis C infection differ in both genders which may help to develop prventive strategies separately. In a descriptive study; The history of theraputic injections (72%) and barber shaves (35%) were also repoted as the most recurrent risk factors among the audlts of Larkana City, Karachi (Shaikh *et al.*, 2009). Almost similar factors were reported and supported by another descriptive study from district Bannu, Khyber-Pakhtunkhwa. This study further reported that majority of the cases were ignorant about the spread or transmission of infection (Shaikh *et al.*, 2009). Keeping in view, Alavian and Aalaei-Andabili (2011) emphasized that a strict health policy in Pakistan should be introduced and more risk factors of hepatitis C infection should also be explored inlcuding history of marital relation. This factor was included in a case control study from Bahawalpur district (Southern Punjab) by Qazi and Ijaz (2011). This was a hospital based study of total 2200 patients including 350 cases and 1850 controls. Out of total patients about 59.1% were males and 40.9% were females. The descriptive and Chi-Square analysis imply that history of marital relation was found positive in 17.1% of cases and even identified as significant (Balasekaran *et al.*, 1999) in Bahawalpur district of Southern Punjab. However, still this study was concluded on the basis of descriptive statistics which need to verify with powerful statistical tools. This gap was tried to fill up with a latest cross sectional study from second largest district, Kech of Balochistan province, Pakistan using univariate and multivariate logistic regression analysis. The crude and adjusted odds ratio of associated risk factors were also computed with their 95% CI. The main objectives of this study was to estimate prevalence of HCV infection in the Kech district of Balochistan, however, some common risk factors were also studied additionally. This would be a first large rural, community based study in province (Balochistan), Pakistan with a sample of (n=2000) patients enrolled randomly from the Kech community population. The overall prevalence of HCV infection was identified as 5.5% but relatively higher rates were determined in male population. In multivariate analysis, the authors found that age>=75 years, being health care worker and Injecting drug use (IDUs) were the independent risk factors of hepatitis C while some important factors such as barber shave, tattooing and sharing of razors reported to be insignificant. Thus authors concluded that although some routs of transmission of HCV infection explained but still at the

population level certain potential risk factors remained un-explained. The risk factor" Injecting drug use (IDUs)" are first time reported in Pakistan although this factor was most often significant and frequent in international studies (Kaldor *et al.*, 1992, Delage *et al.*, 1999, Mishra *et al.*, 2003). This was only because of ease in availability of opium supply from our neighouring country Afghanistan, making injectable drugs readily accessible (Ahmed *et al.*, 2012). Oliveira-Filho *et al.*, (2010) identified almost matching risk factors from Brazil. These were sharing of needles, invasive dental treatment, sharing of razors at barber shops, sharing of cutting material etc. The authors used similar modeling strategy for modeling risk factors.

2.3 Review of Resaerch Gap

In summary, the above literature review illustrates that the risk factors for the transmission of HCV infection may differ considerably within and among the countries. Most of the studies were focused and targeted on specific population or settings, for example, blood donors, hemodialysis patients, injecting drug users, pregnant women and health care workers to identify potential risk factors of hepatitis C infection (Neal et al., 1994, Montagnac et al., 1994, Mele et al., 1994, Delage et al., 1999, Alavian et al., 2003, Zaller et al., 2004, Khan et al., 2008a, Janjua et al., 2010). Only the few studies addressed the general population in order to identify risk factors of hepatitis C with proper study design and statistical analysis (Kim et al., 1996, Miranda et al., 2008, Nguyen et al., 2007, Gheorghe et al., 2010). The situation worsens further when observed for Pakistan specifically. Naturally, this asks for more epidemiological studies in the region on immediate basis. In addition, with reference to Pakistan, most of the studies were reported from Karachi, Sindh province of Pakistan (Akhtar et al., 2004, Neal et al., 1994, Shaikh et al., 2009, Akhtar et al., 2002, Jafri et al., 2006). While a few studies were reported from Punjab province with similar objectives. Again it is alarming that whatever few studies were carried out for Punjab, majority of them used basic descriptive analysis and could not full fill the desired objectives. The most commonly used statistical tools were Fisher's Exact test (Chi-Square) for determining any association between two categorical exposures; t-test to compare the means of two continuous variables, Univariate and multivariate Logistic regression to model and analyze the potential risk factors of hepatitis C infection. It was observed that although some studies employed multivariate logistic regression model but they could not describe model diagnostic checks by adding model adequacy, outlier checking, multicollinearity and sensitivity analysis for better reliable results. This study aims to fill this gap with well-established statistical analysis of risk factors of hepatitis C data in Punjab province with advance or relatively new

statistical approaches besides this traditional logistic regression. These comparative techniques are Artificial Neural Networks and Classification trees which may give us a new direction and better insight of different determinants of hepatitis C in Pakistani's population. The importance and usefulness of these techniques are also presented in the following subsections:-

2.4 Artificial Neural Networks

Logistic regression has been widely used for identifying risk factors and for prediction, but it has some limitations, for example, the handling of interaction effects, missing values and outliers and multicollinearity. Alternative statistical methods such as artificial neural networks (ANN) and classification trees have, in recent years, become popular, not so much to identify risk factors, but as prediction tools (Xue et al., 1996, Sgourakis et al., 2012, Hong et al., 2011, Heydari et al., 2011, Jajoo et al., 2002). Artificial neural networks (ANN) are able to detect hidden complex relationships which may not be captured by conventional regression techniques. ANN models implicitly detect all possible hierarchical interactions which may increase the model performance and may improve predictive accuracy (Hakimpoor *et al.*, 2011). This task remains impractical and laborious using traditional methods. In addition, ANNs are highly flexible since they can handle a mixture of variable types (continuous, nominal or ordinal) in the same analysis, and are generally efficient and do not demand strict distributional assumptions, unlike the LR model.

There are certain limitations of logistic regression including that it does not rank the variables in order of importance (Sarle, 2000) which can be achieved through applying ANN model. ANN model can assess the ranking of individual important variables. However, unlike ANN and CART, logistic regression not only identifies significant factors but also measures the strength of association between the outcome and risk factors, which is greater benefit than ANN and CART. Some authors have compared the ANN model with the LR method in other medical contexts (Tu, 1996), however, the application of ANN model is not so prevalent in epidemiological studies. For example, in one case-control study of myocardial infarction (Vineis *et al.*, 1997), the authors introduced the idea of the relative importance of variables, and suggested that researchers should also compare these two techniques in future studies.

The two methods have been compared in studying risk factors of other diseases; for example, in diabetes, Salmonella typhimurium infections, and coronary artery bypass (Qin et al., 2005, Voss et al., 2002, Gao et al., 2004). Another study of the risk factors of uterine

myomas using ANN model and LR demonstrated similar leading risk factors and on comparison with conventional statistics, ANN perform better, on measures of discrimination and provided powerful alternative to risk factors analysis (wie, 2007). The same conclusion was supported by other studies (Flaherty and Patterson, 2003, Dussol *et al.*, 2007) that on comparison ANN model is not more efficient than an LR model in discriminating the different parameters. Hart & Wyatt (1990) presented a critique that "black box" feature is the key obstacle to the use of neural networks for medical decision making. If prediction is the only objective, then ANN models provide acceptable results, whereas logistic regression could also recognize the effect of factors on the classification (Kazemnejad *et al.*, 2010).

2.5 Classification Trees

Classification Tree methods are very useful alternative of traditional statistical methods and have gained popularity in the last decades to predict the binary outcome from the number of independent variables. Unlike the Artificial Neural Networks (ANN) this is a non-parametric technique and did not require any distributional assumption compared to logistic regression model. Currently, different types of classification trees are available; however, Classification and Regression Tree (CART) method is a common and powerful technique to predict and classify the binary outcome with easy interpretation of visual output tree diagram. This method was first time developed by Breiman (Breiman *et al.*, 1984) and has increasingly been applied in clinical research and prediction purpose(Curran Jr *et al.*, 1993, Falconer *et al.*, 1994, Temkin *et al.*, 1995, Hess *et al.*, 1999) and, to a very lesser extent, in public health and epidemiological studies (Lemon *et al.*, 2003). On the other hand only few studies have been conducted to determining and identifying potential risk factors in different context but not so for hepatitis C infection with the use of Classification Trees. For example, Hasford *et al* (1993) performed both CART and logistic regression analysis to evaluate risk factors of first dose hypertension patients and concluded that CART is an important complement for risk factor identification. A case control study was also presented by Nelson *et al* (1998) to model risk factors of Subarachnoid Hemorrhage patients. In this study authors used the classification tree method and validated by traditional conditional logistic regression to identify risk factors of Hemorrhage patients in a case control data. The results reveal that as a supplemental method, Classification Tree method not only classifies subgroups with varying risks, but also may expose interactions between predictor variables. The idea of this study was further supported by another case control study with application of

Classification Tree analysis for colon cancer in United States (Camp and Slattery, 2002b). Total 4403 patients were enrolled with 2410 controls and 1993 cases. The final Classification tree model suggested that about 61% patients were correctly classified as case or control. The study also concluded that we might be better able to identify risk factors that increase susceptibility to disease by accounting for interactions between risk (Nelson *et al.*, 1998, Kuchibhatla and Fillenbaum, 2002).

Camdeviren *et al* (2005) suggested that although logistic regression is most widely and acceptable method for identification of risk factors but certain issues are still there, for example, handling of missing values, interaction effects and restriction of distributional assumptions etc. In such a situation regression tree method can be a useful and effective approach than traditional statistical methods in epidemiological studies which identify risk factors. Ture *et al* (2005) compared different Classification Tree methods and ANN model with the Logistic Regression model on hypertension case control data and found that CART performed well than others. Similarly, in another study, on comparing Logistic Regression with Classification Tree method almost matching risk factors were determined that manifest the role and application of CT method in risk factors identification (Kitsantas *et al.*, 2006). A concordant conclusion was made by Camdeviren *et al* (2007) while identifying social and demographic risk factors for postpartum depression data. This study reveals that three risk factors were identified by the logistic regression while in case of Classification Tree method those were six. This signifies the importance of Classification Trees that some time we may not capture all possible risk factors with the help of Logistic regression, so that the application of Classification Trees may help out in this regard. In the same context, Thang *et al* (2008) identified risk factors for Malaria combining multivariate and CART analysis to better define the inter-relationships between the different risk factors and constitute a novel analytical approach (Guo *et al.*, 2006). Nagy *et al* (2010) showed that Tree based method can be a new alternative in enlightening risk factors, either used alone or combine with logistic regression model. Another study was performed by Piper *et al* (2011) using Decision Tree (Classification Trees) to identify risk factors for relapse to smoking and results suggested the dynamic importance of method. All aforementioned studies were supported by this latest study (Afonso *et al.*, 2012) and highlighted the significance and application of CART model in identification of risk factors.

2.6 Review of Research Methodology

By reviewing various studies in section 2.1 & 2.2, it was investigated that the most common technique in exploring the potential risk factors of hepatitis C was the multiple logistic regression model. For example, Darwish *et al.*,(1993) identified risk factors of HCV infection in Egyptian blood donors using the same technique. Similarly, Neal *et al.*, (1994) presented a case control study on blood donors in Trent region, UK wherein authors applied univariate and unconditional logistic regression model. Odds ratios and their 95% CIs for each significant risk factor were also computed in the model. Another hospital, case control study from Itly also applied the same technique, in the same context (Mele *et al.*, 1994). However, this study also included socio-demographic factors. A parallel case control study, from Korean population applied the similar technique to meet the same objectives (Kim *et al.*, 1996).

In a similar way, Merle *et al.*, (1999) carried out population based case control study in France; Briggs *et al.*,(2001) in urban veterans, Habib *et al.*,(2001) and Medhat *et al.*,(2002) from the Egypt, Alavian *et al.*,(2002) from Iranian blood transfuion organization, Kim *et al.*, (2002) in Korean population applied the same technique. A study from our neighboring country China was carried out on elderly population of Taiwan by Lin *et al* (2003). Hepburn and Lawitz,(2004) presented a similar study from urban population of Haiti. Moreover, all variables which possessed p-value<0.20 were futher included in the multivariate logistic regression model. Yazdanpanah *et al.*, (2005) designed a case control study among the health-care professionals in five European countries. For statistical analysis, crude and adjusted odds ratios were computed using univariate and multivariate logistic regression. Variables at p-value<0.25 were selected for final selection in multivariate logistic model. Backward step-wise variable selection method was used. Moreover, to check model adequacy, goodness of fit test was also applied. A study from another neighboring country, Iran was also carried out with the same concept in which only univariate analysis was applied and lacking multivariate analysis. Ben-Alaya-Bouafif *et al* (2007), Nguyen *et al* (2007), Wolff *et al* (2008), Liu *et al* (2009), Gheorghe *et al* (2010), He *et al.*,(2011) and Kandeel *et al*.(2012) also applied the multiple logistic regression in these international studies. Few authors from Pakistan also used the identical method for evaluation of risk factors in their epidemiological studies. For example, (Bari *et al.*,(2001), (Akhtar *et al.*, 2002), Shazi and Abbas (2006), Ijaz and Akhter (2007), Ghaffar *et al* (2009). The above references indicate that multiple logistic regression was commonly used statisical technique for identification of

risk factors. However, alternative techniques such as Arififical Neural Networks (ANN) and Classification Tree models have also gained popularity in recent years, not so much to identify risk factors, but as prediction tools (Xue et al., 1996, Sgourakis et al., 2012, Hong et al., 2011, Heydari et al., 2011, Jajoo et al., 2002). These techniques are non-parametric and do not require any distributional assumptions. For example, in a case-control study of myocardial infarction, Vineis *et al*.,(1997), introduced the application of ANN model by suggesting that researchers should also compare ANN model with the logistic regression in their future studies. ANN models could not be applied on hepatitis C risk factors data in any study, but applied in studying risk factors of other diseases, for example, in diabetes, Salmonella typhimurium infections, and coronary artery bypass (Qin et al., 2005, Voss et al., 2002, Gao et al., 2004). In these studies results from ANN model were compared with logistic regression model as well. In another study of risk factors of uterine myomas, ANN model was applied (wie, 2007). This study also provided that ANN model is a power tool to achieve similar objectives.

The importance of Classification Tree models can never be ignored. Hasford *et al* (1993) performed both CART and logistic regression analysis to evaluate risk factors of first dose hypertension patients. The authors concluded that CART model can be a useful tool in risk factors identification. Others studies by Nelson *et al* (1998), Camp and Slattery, (2002b), Ture *et al* (2005), Kitsantas *et al*., (2006), and Camdeviren *et al* (2007) also applied the classification tree methods to evaluate risk factors of some other diseases but not for hepatitis C. Thang *et al* (2008), Nagy *et al* (2010) and Piper *et al* (2011) also emphasized the importance and use of Classification Tree models on risk factors data with or without logistic regression model. The use of these type of models is however, supported by Afonso *et al*.,(2012) in their recent study.

In summary, the above discussion illustrated that Classification Trees and ANN models can also be used besides Logistic Regression model in analyzing the potential risk factors of hepatitis C disease.

2.7 Summary of Major Findings from the Literature

A detail literature regarding HCV epidemiology has been mentioned above which point outs that various studies have now become the part of national and international literature. These studies highlight the importance of issue taken under consideration and reveal that almost every country has now been confronting with this disease. The prevalence

of disease also varies by region and settings of population. The literature showed that the routs or causes of HCV infection are now well recognized at some extent but still it was need to explore its more hidden sources of transmission in different regions of the world. A variety of the risk factors were reported in literature but blood transfusion and injecting drug use were stood prominent and ongoing source of transmission of HCV infection (Serfaty *et al.*, 1993, Comandini *et al.*, 1998, Briggs *et al.*, 2001, Hand and Vasquez, 2005, Souto *et al.*, 2012). But, major risk factors established in developing countries were the surgeries, tattooing, history of injections, dental surgery, hospitalization, body piercing and sharing of needles/razors etc (Kandeel et al., 2012, Awadalla et al., 2011, Kim et al., 2002, Akhtar et al., 2002). In Pakistan's perspective, the most common risk factors were the history of un-safe injection and blood transfusion (Luby *et al.*, 1997). While the poor socio-economic status, low level of education, barber shave, toothbrush sharing and sharing of razors were also the additional possible causes of infection in Pakistan (Bari *et al.*, 2001, Janjua *et al.*, 2005, Shazi and Abbas, 2006). Some studies from international literature also reported that un-safe marital relation may also be the source of transmission but in small percentage (Mendes-Correa et al., 2005, Briggs et al., 2001). Gender-wise comparison of risk factors was lacking in the literature. Moreover, risk factors among the urban/rural settings of patients' could also not be studied previously which are now accompanied in this study.

Chapter 3

RESEARCH METHODOLOGY

Research methodology is a structured procedure of conducting research to achieve the desired objectives of the study. This usually includes study design, sample size determination, data collection procedure, statistical or theoretical frame work and data analysis scheme. Conversely, the ultimate objectives of defining the research methodology are to cover the procedures followed to analyze the collected data and then to interpret the results. This study aims to statistically model the risk factors of hepatitis C in the province of Punjab, Pakistan, using three different statistical approaches (logistic regression (gold standard), artificial neural networks and classification tree methods. The theoretical frame work for these methods is explained in the subsequent sections. However, at first, study design, sampling and sample size determination, Instrument of the study, pre-testing, coding and illustration of variables are discussed.

3.1 Statistical Terms (Online Statistical Dictionary)

Some useful statistical terminologies which can help to understand the methodology concepts are also added below before describing detailed part of the research methodology.

- **Statistical Modeling:** In statistical modeling a relationship between the predictors and the outcome variables is established through an estimated mathematical model. A model based on significant predictors is fitted to the sample which intends to predict dependent variable by estimating parameters in the population and inference is drawn about the population by testing of hypothesis and construction of confidence intervals.
- **Data Mining Methods:** These are the statistical techniques which help in exploring a large set of variables or data and discover the hidden relationships between variables very sophistically.
- **P-value:** Probability value is the probability of obtaining, by chance alone, results at least as extreme as those obtained equivalently.
- **Observational Study:** Observational studies include case-control and cohort (prospective and retrospectively) studies wherein current behavior is observed without performing an experiment or intervention.

- **Case Control Study:** A case-control study is the type of epidemiological study which involves in grouping subjects with the condition being investigated ‘cases (diseased subjects) and compared with ‘controls (disease free subjects). These types of studies help to identify whether the cases are more or less likely than controls to have had particular risk factors.
- **Cross Sectional Study:** In this type of study information pertaining to particular disease is collected from the individuals in a population during fixed or defined time period.
- **Consecutive Sample:** A **s**ample in which the subjects are selected on the basis of first come first chosen. All subjects who meet the eligibility criteria should be included as they are found.
- **Odds ratio:** It is the commonly known measure of association which measures the association of an exposure with the disease. It is the necessary part of the case-control studies and help in evaluating the risk of the disease based on particular risk factor. It is the ratio of two odds which yields an approximate value for the relative risk of the exposure.
- **Multicollinearity:** Multicollinearity is the statistical term which indicates strong correlations among the explanatory variables when regression models are constructed. Presence of multicollinearity often results in large standard errors (SE) and insignificant coefficients.
- **Outlier:** A value which is far away from rest of the data is known as an extreme score or outlier in statistical terms. It generally occurs due to an error in measurement or data entry or recording.
- **ROC Curve:** It is a useful graph which helps in determining the ability of the fitted regression model and discriminates the individuals who have or have not a particular disease. This graph is generally plotted between sensitivity and specificity measures.
- **Sensitivity:** Sensitivity measures the proportion of actual positives subjects which are correctly identified by the fitted model or a diagnostic test.
- **Specificity:** It measures the proportion of actual negatives individuals which are correctly identified through the fitted model or a diagnostic test.

3.2 Study Design

The selection of a proper study design is often of decisive importance for research planning (Röhrig *et al.*, 2009). There are different study designs available in medical research; however, one which helps to meet the desired study objectives has to select by the

researchers. This study was a hospital-based case control study, conducted for investigating potential risk factors of HCV infection. Case control study is a type of observational study, generally designed for determining the relationship of an exposure with an outcome i.e. disease (Lewallen and Courtright, 1998). These types of studies are easy, quick and inexpensive (Röhrig *et al.*, 2009). Characteristics of individuals who have a specific disease (cases) are compared with those who are disease free (controls). It is advantageous to note that controls should be selected from the same population from which cases are being chosen (Khan et al., 2008b). Various studies with a view to determine the risk factors of HCV infection have used the similar case control study design. For example, Neal *et al.*, (1994), Mele *et al.*, (1994), Chiaramonte *et al.*,(1996), Comandini *et al.*, (1998), Briggs *et al.*,(2001) and Karmochkine *et al.*, Karmochkine *et al.*,(2006a) conducted similar hospital based case control studies for risk factors identification.

3.3 Sample Size Estimation and Data Collection Procedure

Pakistan is the sixth most populous country in the world, with a population of about 180 million people. Among its four provinces, Punjab is the largest, comprising about 60% of the country's total population. Administratively, Punjab has been divided into 9 divisions (Lahore, Gujranwala, Rawalpindi, Faisalabad, Multan, Sargodha, Bahawalpur, DG Khan, Sahiwal) which contain a large public sector Divisional Head Quarter Hospital managed by government of the Punjab **Table 3.1**. Here free treatment for HCV positive patients is available, sponsored by the Government of the Punjab so that men/women, all social strata, urban/rural patients belonging to adjacent districts and communities have opportunity for free diagnosis and treatment. This hospital based study, now carried out for investigating potential risk factors of HCV infection in the Punjab province.

As a pre-step, permission was sought from the hospital administration for in-person dialogue with patients for data collection. At first, the researcher briefed the physician about the study objectives for his support in identification of the cases and controls. The physicians supported the researcher to meet and interview the patients in medical in-patient, out-patient and hepatitis clinics established within the hospitals. By following this procedure, a cross-sectional survey of cases and controls were made from each Divisional Head Quarter hospital throughout the province of Punjab. Consecutive sampling method was used for selection of cases and controls (Ghaffar et al., 2009, Zeuzem et al., 1996, Chlabicz et al., 2006). The same consecutive sampling method was opted in collection of all data using a structured questionnaire which included demographic, socio-economic, family history and clinical

factors. In Pakistan, there is no data collection system available for recording routine risk factors of hepatitis C patients (Qidwai *et al.*, 2010) Therefore, the researcher (MG) travelled throughout Punjab province to interview the patients and collect the data from each divisional hospital during a six month period.

A sample size for this retrospective, unmatched case-control study was estimated using (Open-Epi) version 2.3.1 online software. The results were presented using Fleiss (1981) formula. The estimated sample was 1396 assuming case to control ratio 1:1 (Kaldor *et al.*, 1992, Mendes-Correa *et al.*, 2005), with a probability of a type 1 error set at 5% and statistical power of 80% (Onyango *et al.*, 2013). Mendes-Correa et al.,(2005) also estimated the sample size for their case control study to evaluate risk factors of HCV infection among the HIV patients in Brazil with similar power (80%) and confidence interval (95% CI). For convenience, the overall estimated sample size was rounded off to 1400. A double sample size was allocated to the large divisions of Punjab as compared to small divisions. The use of this Fleiss formula in epidemiological was explained by Kasiulevičius et al.,(2006). Whilst a latest study by Onyango et al.,(2013) also showed its application in case control study for risk factors identification of sever pneumonia in children of Western Kenya. Pre-requisite estimated proportions of some established risk factors against cases and controls were taken from previously published literature and incorporated in the Fleiss formula. The complete distribution of total sample size is described in **Table 3.1**. Moreover, prior to getting the thorough information on prescribed proforma from each patient, the purpose of the study was also explained to every case and control and proper time was given to every eligible respondent for interviewing and filling each questionnaire to circumvent from recall bias. The predetermined sample was collected by different consecutive visits of each selected hospital against each corresponding division of the province Punjab and the process of collection remained continue until the entire required sample was covered. Consequently, a total sample of 1400 patients as estimated (Including; 700 cases and 700 controls) with equal proportion was collected by surveying all selected hospitals during the period of 6 months. For detailed illustration, Fliess formula is explained below:-

$$n_1 = \frac{[z_{\alpha/2}\sqrt{(r+1)\overline{pq}} + z_{1-\beta}\sqrt{rp_1q_1 + p_2q_2}]^2}{r(p_1 - p_2)^2}$$

$$n_2 = rn_1$$

$$n_{1cc} = \frac{n_1}{4}[1+\sqrt{1+\frac{2(r+1)}{n_1 r(p_2 - p_1)}}]$$

$$n_{2cc} = rn_{1cc}$$

n_1 =number of cases

n_2 =number of controls

$z_{\alpha/2}$ = Standard normal deviate for two tailed test based on alpha level (relates to the confidence interval)

$z_{1-\beta}$ = Standard normal deviate for two tailed test based on beta level (related to the power level)

r_1= ratio of controls to cases

p_1=proportion of cases with exposure and q_1= 1- p_1

p_2=proportion of controls with exposure and q_2= 1- p_2

Table 3.1: Allocation of Sample Size in Each Selected Hospital

Sr .No.	Division	Hospital Name	Total	Cases	Controls
1	Lahore	Services hospital	**200**	100	100
2	Gujranwala	DHQ Hospital	**200**	100	100
3	Rawalpindi	Holy family	**200**	100	100
4	Faisalabad	Allied hospital	**200**	100	100
5	Multan	Nishtar hospital	**200**	100	100
6	Sargodha	DHQ hospital	**100**	50	50
7	Bahawalpur	Victoria hospital	**100**	50	50
8	DG Khan	DHQ hospital	**100**	50	50
9	Sahiwal	Civil hospital	**100**	50	50
Total			**1400**	700	700

3.4 Cases and Control Selection

A consecutive sample of 700 cases and 700 controls patients were identified from the outpatients, inpatients and hepatitis clinics (where applicable) out of each selected hospital. The cases were all those patients who were subsequently determined to be HCV positive from blood samples using a routine ELISA method, whilst controls were all negative. The controls, in addition, were HBsAg negative and were not clinically jaundiced. The patients of every age group and gender were included in the study with the only exclusion criterion being individuals who were cognitively impaired. The cases were likely to visit the health care clinic for consideration, initiation or follow-up of drug treatment for HCV infection, whereas controls were chosen from medical clinic visitors who were acutely ill, with signs and symptoms such as fever, nausea, vomiting, body pain, poor appetite and were looking for a medical diagnosis as well as treatment at the same hospital. Both cases and controls possessed no evidence of serious liver synthetic dysfunction, such as high International Normalized Ratio (INR), low albumin or low platelet count. In other words, both cases and controls had no clinical or biomedical evidence of advanced liver disease.

3.5 Development of Questionnaire

In this study, the questionnaire was designed by the researcher in consultation with his supervisor and a Gastroenterologist. Local as well as published knowledge was employed to make the questionnaire culturally acceptable. The questionnaire was included information about socio-demographic characteristics of patients, family history, patients' behavioral characteristics and medical history related factors. It was administered by a researcher to both cases and controls in each hospital.

3.6 Pre-Testing

Pre-testing is a method to appraise the validity of the questionnaire and to see if it works adequately in the field or not. This may also help to improve the questionnaire by including or excluding some variables and avoiding problems in upcoming thorough survey. In this study, about 40 eligible patients (including 20 cases and 20 controls) were interviewed by the researcher from the Services Hospital, Lahore who were visiting hepatitis clinic for diagnosis and treatment of hepatitis C. After filling out 40 questionnaires, an analysis was made and few questions which were not responded were deleted, for example, "*Injections received from multiuse vials*". This is a reported risk factor in Pakistan's perspective that majority of the un-registered medical practitioner used to have injections from multiuse vials to the patients but

during pre-testing it was observed that people were often unclear about its history. The main reason was that the un-registered medical practitioners conceal this fact from the patients. Similarly, another suspected risk factor *i.e* "Extramarital marital relationship" was still considered in the overall subsequent survey and history of marital relationship was asked only from male patients. This was also a reported risk factor in other countries, such as Canada, Barazil, Iran and Thailand but could not be studied in Pakistan because of societal constraints. It was expected that people might be reluctant to answer this question but unexpectedly they responded to it well during the pre-testing. Therefore, it was decided to include this risk factor in the subsequent survey. Alavian and Aalaei-Andabili (2011) also revealed to the importance of this factor and suggested that history of marital relation should also be evaluated in Pakistani population.

3.7 Interview Bias in Data Collection

Data collection is the most pertinent phase of any study or research. It provides the basis for conducting an authenticated research but there may be certain issues like interview bias, which need to be addressed efficiently. Interview bias is simply the personal tendency of the interviewer towards some specific questions or otherwise. There are different types of interviews which enable us to collect data for conducting studies, for example, structured interview, semi-structured interviews, and unstructured interviews. This study is conducted through the structured interview which is relatively a secured method as compared to others. This is due to the fact that each member of the sample group is asked the same structured questions, and hence the chances of interviewee bias minimized. During data collection, a due care was given to avoid interview bias by implementing following steps:-

a) All the sampled patients were interviewed in person by the researcher.
b) All the selected patients were properly briefed about the purpose/usefulness of the study before conducting the interview.
c) A proper time was allocated for interview on each patient to avoid recall bias.
d) Doubts about the study, if any, were properly clarified.
e) Confidentiality of record/information was ensured.
f) Questions were asked in simple understandable language.

Overall, this study avoided interview bias best possibly and authenticity of the data was maintained.

Table 3.2: Comparison of LR, ANN and CART models Characteristics

Parameter	Logistic Regression	Artificial Neural Networks	Classification Trees
Model Building	It has more comprehensive statistical knowledge and its background.	Less statistical knowledge and its background.	Entails less statistical knowledge and its background.
Distributional checks	It is parametric technique and requires variables distributional checks.	It is a non-parametric technique and does not require any distributional checks.	It is a non-parametric technique and does not require any checking of distributional assumptions.
Diagram	No graphical presentation is available to describe pertinent factors.	Diagram exists but not interpretable.	Results are easily understandable and interpreted visually through a tree diagram.
Regression Coefficients	Regression coefficients are properly estimated through (MLE) method which helps in interpreting the results in terms of odds ratio.	Regression Coefficients do exist but cannot be interpreted like LR model.	Regression Coefficients cannot be calculated.
Goodness of Fit	For goodness of fit, proper statistical tests are available, e.g. Hosmer & Lemeshow test etc.	No proper goodness of fit criteria available, however, it can be done with the help of ROC curve analysis, examining the correct classification of outcome variable, measuring of sensitivity and specificity.	No proper goodness of fit criteria available, however, it can be done with the help of ROC curve analysis, examining the correct classification of outcome variable, measuring of sensitivity and specificity.
Ability to detect complex relationships	Fails to discover complex relationships between the outcome and the predictors unless researcher	Automatically detect hidden complex relationships between outcome and the predictors.	Automatically detect the hidden complex relationships between outcome and the predictors.

	does not define in advance.		
Ability to detect interaction effects	All possible interactions are difficult to manage due to complexity of the model. So, only the selected pair of interactions are acceptable.	ANN model has ability to detect interactions automatically and even higher order interaction effects are considered in the model. However, it is difficult to learn about the exact combinations of interaction effects because only the network knows about these interactions inside and can never be seen apparently.	Unlike LR and ANN models, these models automatically detect multilevel interactions very efficiently in a tree diagram. These are observable and easily interpretable. So, these models are powerful tool in assessing physical and social interactions which could not be accompanied in the traditional approaches.
Missing values	Missing values should have to discard preferably; otherwise, LR model can never be run. This deletion ultimately affects the sample size. Additionally, the replacement of missing values augments the volume of analysis.	ANN model can be run in the presence of missing values without reducing sample size.	Missing values can be handled very efficiently and without losing sampling size. CART simply uses surrogate splitters to handle missing values.
Selection of variables and their interpretation	Significant variables are easily identifiable and interpretable in a straight forward manner.	Difficult to choose pertinent factors **"Black Box"**.	Important variables are easily explored among the large number of variables and their interpretation is also straight forward.
Ranking of variables	Ranking of significant variables is not achievable.	Ranking of variables is made by calculating Normalized importance of each variable. A variable having its largest value will be ranked first and so forth.	Important variables are selected in repeated levels and arranged according to their level of importance. The most pertinent factor would be selected at first level and this order of arranging variables comes down with increasing number of tree levels.

Over fitting	Over fitting is less of a concern	Over fitting is a great concerned but can be handled by partitioning the data into training and testing sets as discussed in this study.	Over fitting is not so problematic.
Adoptability	Fitted model can easily shared with other researchers	Fitted model is not easily shared	Fitted model is not easily shared.
Calculation of confidence intervals	Easy	Difficult	Not calculated

3.8 Description and Coding Scheme of Variables

Information was gathered using a structured questionnaire which contained socio-demographic (age, gender, marital status, level of education, family status, geographic location, monthly household income, paternal or maternal education) information. Socioeconomic status was classified according to monthly household income, and categorized into three further groups. A list of clinical risk factors included: history of injection, parenteral exposure to blood or blood products, dental surgeries, previous hospitalizations, accidental needle stick, surgical/medical procedures, minor surgery by a barber (for example: circumcision, ingrown toenail surgery or drainage of abscesses over the time), personal or family history of jaundice/ hepatitis, and history of cuts (from a blade or knife). Other factors included a history of body piercing, tattooing, sharing of toothbrushes or miswak (a natural twig used as a traditional toothbrush) or communal nail cutter, history of imprisonment, use of injecting drugs, or local or foreign migration was also included in the study. Moreover, the details of each variable or risk factor are given in **Figure 3.1** i.e. conceptual framework of data variables.

Figure 3.1: Conceptual Framework of Data Variables

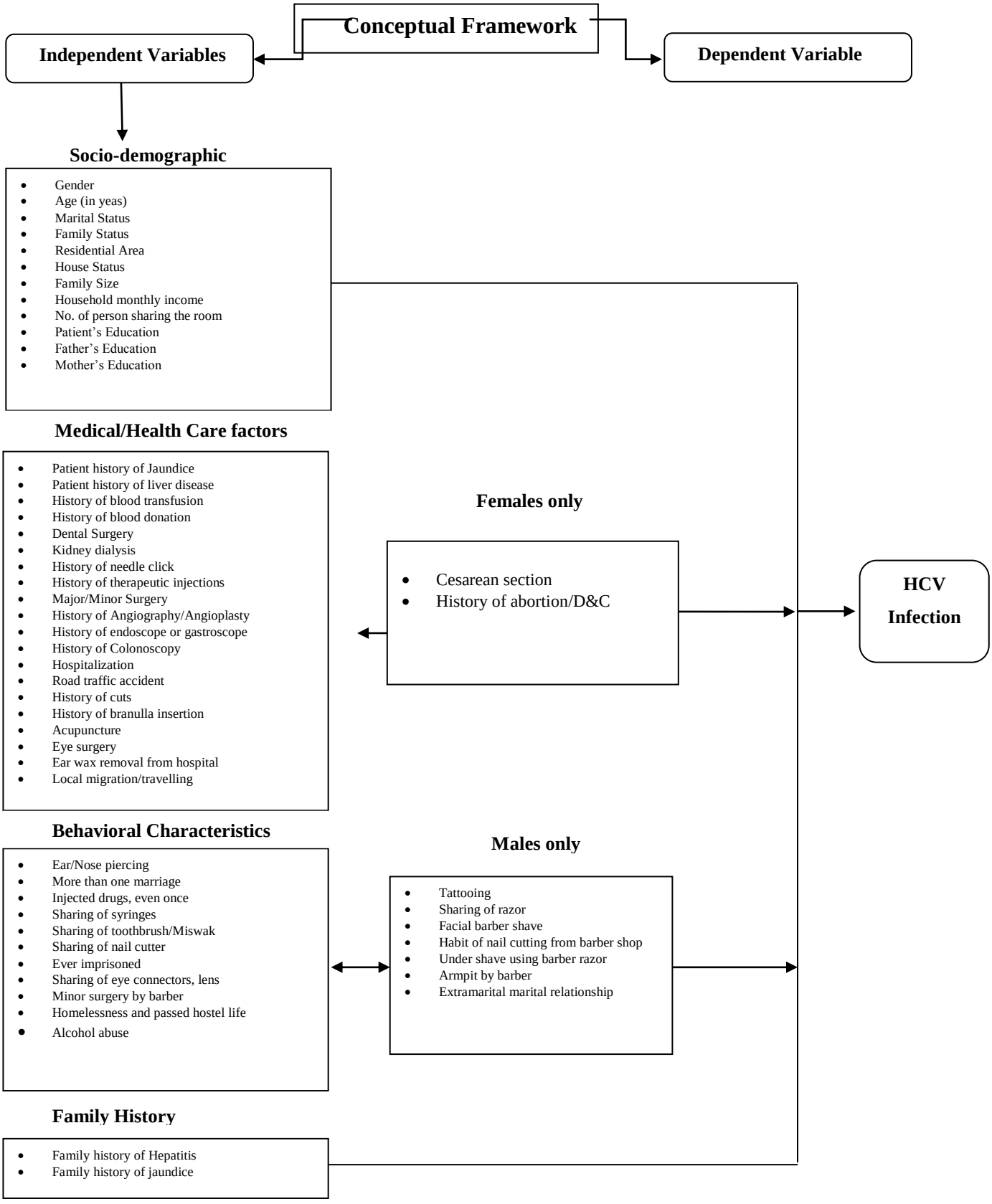


Table 3.3: Coding Scheme of All variables under study

Sr.No.	Variable	Category	Coding Scheme	Reference Category	Description of variables
1.	Gender	Male	1	First	Male verses female patients
		Female	2		
2.	Patient's age group	<=34	0	First	Initially, patient's age was taken as continuous variable but later on divided into two groups taking cut point at mean age.
		>34	1		
3.	Marital Status	Never married	1	First	Married patients verses un-married
		Ever married	2		
4.	Family status	Single	1	First	People of Pakistan live in nuclear as well as joint families. Joint family is an extended family system.
		Joint	2		
5.	Residential area	Urban	1	First	According to the residential area of the patients. The patients were categorized into the urban and rural areas.
		Rural	2		
6.	House status	Owned	1	First	Patients who had lived or living in rented houses verses they had their house in an ownership.
		Rented	2		
7.	Family size	-	-	-	No. of persons living in a house
8.	No.of person sharing the room	-	-	-	No. of persons sharing the room in a same house or flat.
9.	Patient's Education	None	0	Last	Patient's education was divided into four groups according to years of formal education they have received. First group indicates no or zero education or simply patients "illiterate".
		1-5	1		
		6-10	2		
		11-16	3		
10.	Income group (in PKR)	<=5000	1	Last	Household income was divided into four groups.
		5000-15000	2		
		15000-25000	3		
		>25000	4		
11.	Father's Education	No	0	First	Patients were inquired about their father's education (Yes/No)
		Yes	1		
12.	Mother's Education	No	0	First	Patients were inquired about their mother's education (Yes/No)
		Yes	1		
13.	Family history of Hepatitis c	No	0	First	Family history of hepatitis C was inquired from the patients only about the first order relatives.
		Yes	1		
14.	Patient history of Jaundice	No	0	First	Patients had or had not history of jaundice (Yes/No)
		Yes	1		
15.	Family history of jaundice	No	0	First	Family history of jaundice (Yes/No)
		Yes	1		
16.	Family history of liver disease	No	0	First	
		Yes	1		
17.	History of blood transfusion	No	0	First	
		Yes	1		
18.	History of blood donation	No	0	First	
		Yes	1		

19.	Dental surgery	No	0	First	
20.		Yes	1		
21.	Tattooing	No	0	First	
		Yes	1		
22.	Ear/nose piercing	No	0	First	
		Yes	1		
23.	More than one marriages	No	0	First	
		Yes	1		
24.	Injected drugs, even once	No	0	First	
		Yes	1		
25.	History of accidental needle stick	No	0	First	
		Yes	1		
26.	Sharing of syringes	No	0	First	
		Yes	1		
27.	Sharing of tooth brush/Miswak	No	0	First	Sharing of toothbrushes or miswak (a twig used as a traditional toothbrush)
		Yes	1		
28.	Communal nail cutter	No	0	First	Most of the persons in a household share a single nail cutter for nails cutting purpose.
		Yes	1		
29.	Major/Minor Surgery	No	0	First	Surgical/medical procedures involve cut or stitching.
		Yes	1		
30.	History of Angiography/Angioplasty	No	0	First	Some patients had the history of angiography or angioplasty. It is reported that proper sterilization of instruments is not ascertained in hospitals during the procedure, so the HCV may transfer to some other patients.
		Yes	1		
31.	History of endoscope or gastro-scope	No	0	First	It was also a reported risk factor of HCV. Infection may transfer, if instruments shared or not properly sterilized.
		Yes	1		
32.	Road traffic accident	No	0	First	
		Yes	1		
33.	History of cuts	No	0	First	History of cuts (from a blade or knife)
		Yes	1		
34.	Hospitalization	No	0	First	
		Yes	1		
35.	History of injections or intravenous drips	No	0	First	
		Yes	1		
36.	History of branula/cannula insertion	No	0	First	
		Yes	1		
37.	Acupuncture	No	0	First	
		Yes	1		
38.	Ever imprisoned?	No	0	First	
		Yes	1		
39.	Local migration	No	0	First	
		Yes	1		
40.	Foreign migration	No	0	First	
		Yes	1		
41.	Sharing of eye lenses	No	0	First	
		Yes	1		
42.	Minor surgery by barber	No	0	First	In Pakistan, it is reported that minor surgery is performed by the barber (e.g. circumcision, ingrown toenail or drainage of abscesses over the time)

43.		Yes	1		
44.	Eye surgery	No	0	First	
		Yes	1		
45.	Ear wax removal from the hospital	No	0	First	It is normal practice in government hospitals that medical practitioners use same instruments on several patients without sterilization or sterilization is not properly ascertained.
		Yes	1		
46.	Homelessness and hostel life	No	0	First	
		Yes	1		
47.	Alcohol abuse	No	0	First	
		Yes	1		
48.	History of Cesarean section	No	0	First	
		Yes	1		
49.	History of abortion/D&C	No	0	First	
		Yes	1		
50.	Place of delivery	Qualified doctor	0	First	
		LHV/Mid wife	1		
51.	Tattooing	No	0	First	
		Yes	1		
52.	Sharing of razor	No	0	First	
		Yes	1		
53.	Facial shave from barber	No	0	First	
		Yes	1		
54.	Habit of nail cutting from barber	No	0	First	
		Yes	1		
55.	Removal of unwanted hairs from barber shops	No	0	First	
		Yes	1		
56.	Armpit from barber	No	0	First	
		Yes	1		
57.	Extra-marital marital relationship	No	0	First	
		Yes	1		

3.9 Statistical Analysis

All the information was gathered on a structured questionnaire consisting of different variables and these variables were pre-coded for the computer analysis. The data was analyzed descriptively and analytically using IBM SPSS version 19.0. The Classification Tree Model was performed using its Decision Trees module. In descriptive analysis mean ± standard deviation for continuous variables and frequency and percentages for categorical variables were computed. While in analytical analysis, initially, Chi-Square analysis, t-tests (where applicable) and univariate odds ratios from logistic regression were evaluated to shortlist the influential factors. All variables which possessed $p<0.20$ in univariate analysis were entered into multivariate analysis for final selection of pertinent risk factors (Hosmer and Lemeshow, 2000, Mendes-Correa *et al.*, 2005). All *p*-values were 2-tailed. Mendes-Correa *et al* (2005) also performed a case control study with the aim to evaluate potential risk factors of HCV infection in Brazil and adopted similar strategy to build the multivariate logistic regression model. Various other studies in literature also employed the similar method to model and analyze the risk factors of HCV on different sets of data (Yazdanpanah *et al.*, 2005, Ben Alaya Bouafif *et al.*, 2007, Nguyen *et al.*, 2007, Nokhodian *et al.*, 2012). However, identification of outliers' logistic regression modeling was lacking in epidemiological studies which have now been added in a recent study. Moreover, the results from ANN and Classification models are discussed in Chapter No.5.

3.10 Statistical Tests for Measuring the Association

Measuring of an association between outcome and various exposures is an important task of epidemiological studies which helps to establish significant relationship between outcome and different exposures. For this purpose, Chi-square and Fisher's Exact tests are most commonly employed and discussed in the following subsections.

3.10.1 Chi-Square test of Association

Chi-square (χ^2) test is an important nonparametric statistical test most widely used in epidemiological studies to compare frequencies or proportions in two categorical variables (Zibran, 2007). When the research hypothesis is implicit in terms of independence, the test is referred to as Chi-Square test for independence and helps us to test whether one of the variables is significantly associated with the other one (Dawson and Trapp, 2004). The test serves both as a "goodness-of-fit" and as a test for measuring association across two or more dimensions (Howell, 2009). In Chi-Square test both the predictor and criterion variables can

be evaluated on any measurement scale (nominal, ordinal, ratio, or interval). The actual Chi-Square test is generally famous as Pearson's chi-square since it was developed by Karl Pearson in the earlier 1900s. P-value is taken as one sided and computed with Chi-Square statistics which explains that our hypothesis is going to be rejected or accepted. A p-value less than 0.05 indicates that the null hypothesis of no association would be rejected. On the other hand, a significant association lies between the two categorical variables. But it should be understandable that the presence of significant association does not mean that the relationship is due to the cause and effect. The Pearson Chi-Square distribution for testing null (Ho: o_{ij}= e_{ij}) hypothesis is expressed as:

$$\chi^2_{(df)} = \sum \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Where o_{ij} and e_{ij} are the observed as well as expected frequencies; df is the degree of freedom and df=(r-1)(c-1). When all $o_{ij} = e_{ij}$ then value of test statistic (χ^2) have zero value while χ^2 ranges from 0 to ∞. Nokhodian *et al.*, (2012) used this test in their bivariate analysis to determine association of risk factors with the disease outcome.

3.10.2 Fisher's Exact Test

The Fisher's exact is considered asymptotically equivalent to Pearson's chi-square (Camilli, 1995) and is test of choice for a 2x2 contingency table when the number of counts in a table are small. Generally, if expected frequency is found less than 5 in any cell of 2x2 cross table then it is a more appropriate to consult a different statistics such as Fisher's Exact Test instead of Pearson's Chi-Square (Hatcher and Institute, 2003). This test was proposed by the distinguished British statistician Fisher in 1934 and different authors in literature have used Chi-Square and Fisher's Exact Tests in epidemiological studies to establish the relationship between exposure and disease status. For example, Darwish *et al.*,(1993), Merle *et al.*, (1999) and Luby *et al.*, (1997) and Nokhodian *et al.*,(2012) used this test in their univariate analysis. Another cross sectional study by Awadalla *et al.*, (2011) applied Chi-square test to find important risk factors of hepatitis C among the Egyptian blood donors.

3.11 Odds and the Odds Ratio

In the recent medical literature odds ratio have been extensively used in epidemiological studies as a measure of association between a disease and a risk factor (Kahn and Sempos, 1989). Odds ratios are the fundamental choice in case-control studies

because likelihood of diseased (cases) patients can be compared with the healthy (controls) patients easily in these studies. It is a measure of the ratio of odds that a case patient was exposed to a particular risk factor to the odds that a control was exposed to that risk factor.

The odds of an event can be defined as the ratio of the probability of a success to the probability of a failure. For example, the odds of an event **A** is Odds (A)= $\frac{Prob(A)}{1-Prob(A)}=\frac{p}{1-p}$

Similarly, on comparing two sets of binary data, an odds ratio is the relative measure of the odds of a success in one set relative to that in the other. Suppose that p_1 and p_2 are success probabilities for two sets of data, the odds of a success in the i[th] set are $p_i \big/ (1-p_i)$, $i=1,2$

and the odds ratio is expressed as $\boldsymbol{\psi = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_2}{(1-p_2)}}}$ and its value is always ranges between 0 to ∞. The odds ratio actually measures the degree or strength of the association between exposure and disease. An odds ratio of 1 occurs when the odds, and hence the proportions, are the same in the two groups (Kirkwood and Sterne, 2003). A zero value suggests that no association at all. Similarly, OR greater than 1 indicates positive association of particular exposure with the disease and vice versa (Kahn and Sempos, 1989). It should be noted that odds ratio is an estimation of risk not a true risk value.

The following 2x2 table enable us to explain the odds ratio in a simple way.

Risk	Disease Status		Total
	Cases (diseased)	Controls(disease-free)	
Exposed	A	b	*a+b*
Non-Exposed	C	d	*a+d*
Total	*a+c*	*a+d*	*a+b+c+d*

$$\hat{\psi} = \frac{\text{Odds in exposed group}}{\text{odds in unexposed group}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

It is known as cross-product ratio of the 2x2 table.

To construct a CI for the true odds ratio, the logarithm of the estimated odds ratio provides better approximation of normal distribution than the odds ratio itself (Collett, 2002). The approximate S.E of the log odds ratio, $\log \hat{\psi}$, can be expressed as:-

$$se(\log \hat{\psi}) \approx \sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$$

An approximate 100(1-α)% CI for log ψ has limits $\log \hat{\psi} \pm z_{\alpha/2} se(\log \hat{\psi})$, where $z_{\alpha/2}$ is the upper $(100\alpha/2)\%$ point of the standard normal distribution (Collett, 2002). Many researchers have calculated odds ratio in their case control studies. For example, Zaller *et al*., (2004), (Hajiani et al., 2006a) and Zakizad (2009) calculate the odds ratio with their 95% CIs in univariate and multivariate analysis.

3.12 Importance of Statistical Modeling

An understanding of what statistical method is suitable according to the type of data is no doubt decisive, "but it is also important to realize that different statistical methods have much in common, so that an understanding of one method helps in understanding of others" (Kirkwood and Sterne, 2003). The basic aim of modeling is to construct a mathematical demonstration of relationships between a response and a number of independent variables. Regression methods have become a fundamental part of any data analysis with several techniques such as linear regression, discriminant analysis and logistic regression to establish a relationship between response and explanatory variables. In case of binary response (e.g. the presence or absence of a disease) linear regression model cannot be applied since the normality assumption is violated and interpretation of model in terms of predicted values become difficult for dummy or indicator response variable (Petrie and Sabin, 2009). In this situation the most widely and acceptable technique is the logistic regression that has consistently been employed and remained popular in the last decades as a gold standard in modeling binary response with a combination of continuous as well as categorical independent variables (Hosmer and Lemeshow, 2000, Tanwandee et al., 2006, Sandhu et al., 1999). The major area of application of the logistic regression model is to analyze the data from epidemiological studies including cohort studies and both matched/unmatched case-control studies (Collett, 2002, Balasekaran et al., 1999, Briggs et al., 2001). The logistic regression model enables to determine which explanatory variables are pertinent and affect the binary outcome significantly. Moreover, predictive strength of variables can also be determined accordingly (Petrie and Sabin, 2009). Other statistical techniques have also been

the part of literature which compared the results of logistic regression as well. These are Artificial Neural Networks and Classification Trees models (Raghavendra and Srivatsa, 2011, Matsui et al., 2002, Hasford et al., 1993) and discussed in Chapter No.5.

3.13 The Logistic Regression Model

In this study the main objectives are to establish the relationship of different risk factors of HCV infection with the disease outcome. Logistic regression is a mathematical modeling approach that can be used in describing the relationship of several risk factors with binary or dichotomous outcome (Lin et al., 2003, Khan et al., 2008a, Delage et al., 1999, Slinker and Glantz, 2008). The logistic regression model is a part of a class of models known as ***generalized linear models***, introduced by Nelder and Wedderburn (1972) for modeling categorical data. These models are actually the extension of linear models for modeling binary response variable because linear models are not fitting in such situation (Molenberghs, 2003, Nelder and Wedderburn, 1972). For the categorical response, a linear relationship is established between the response and independent variables through some possible transformations, known as ***link function***, so that fitted probabilities could be ensured between 0 and 1. The most commonly described link functions are; logit, probit and complementary log-log (Almeida *et al.*, 2001). Of these, logit transformation is appropriate for the analysis of data in case-control studies particularly because of its ease of interpretation in terms of the log of odds of a success (Kandeel et al., 2012, Ghaffar et al., 2009).

As mentioned above, the response variable in logistic regression is usually dichotomous also known as binary response. It assumes value 1 with a probability of success ***p*** or the value 0 with probability of failure **1-*p***. In case of more than two response categories, the logistic regression model becomes multinomial or plytomous. The univariate and multivariate models of binary logistic regression are given below:-

Suppose that there are **n** binomial observations y_i/n_i, for i=1,2,…,n, and $E(Y_i)=n_ip_i$ where p_i is the corresponding response probability. Thus, the univariate case of linear LR model can be represented as:-

$$\text{logit}\left(p_i\right)=\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta\, x_i \qquad \textbf{3.1}$$

Where, logistic transformation or logit of a success probability $\boldsymbol{p}$ is $\log\{(p/1-p)\}$. After some re-arrangements, the logistic regression model for one explanatory variable may be expressed as:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} \qquad \textbf{3.2}$$

Now the case of multiple logistic regression is described below where generalization of more than one explanatory variables is accounted for and referred to as the "multivariable case".

$$logit(p_j) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}$$

On some re-arrangement

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1+\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})} \qquad \textbf{3.3}$$

A sigmoid relationship is appeared between the response probability and the independent variable weight. The weight of the parameter β determines how the sigmoid curve changes (Agresti, 2007, Hosmer and Lemeshow, 2000). Different authors have described the application of logistic regression model in their epidemiological studies in literature (Qureshi et al., 2009, Wolff et al., 2008, Delage et al., 1999) etc.

3.13.1 Fitting of the Logistic Regression Model

To fit a logistic regression to a given data, the foremost step is the estimation of k+1, unknown parameters, $\beta_0, \beta_1, \ldots, \beta_k$. For this purpose, the most readily used method is maximum likelihood estimation method, whilst the method of least square is generally employed in fitting linear regression (Menard, 2002). The estimation of the coefficients in the multiple logistic regression model and testing for their significance is almost similar as the univariate model. The likelihood function is as under:

$$\boldsymbol{L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}} \qquad \textbf{3.4}$$

The likelihood depends on p_i, the unknown success probabilities which sequentially depend on the $\beta's$ through equation **3.3**. Hence the likelihood function can be deemed as function of β and the ultimate purpose of this function is to reach on those values, $\hat{\beta_1}, \hat{\beta_2}, \ldots, \hat{\beta_k}$ which maximizes the $L(\beta)$ or $\log L(\beta)$. The next important step is the evaluation of derivatives of this likelihood function with respect to the $k+1$ unknown parameters and is expressed as:

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} n_i x_{ji} e^{\eta_i} (1 + e^{\eta_i})^{-1}$$

$$j = o, 1, 2, \ldots, k \qquad \textbf{3.5}$$

After evaluating these derivatives at $\hat{\beta}$ for the unknown parameters $\hat{\beta_j}$ and equating to zero, a set of $k+1$ non-linear equations be obtained and solved numerically. For logistic regression, these equations as being the non-linear in nature required special method for their solution including *iterative weighted least squares* method (Collett, 2002). These iteration are required sufficient calculation and manually not practicable. However, with the use of computer software, this can be done easily (Hosmer and Lemeshow, 2000).

Once the $\hat{\beta}$ has been estimated, the relationship between predicted probability and values $x_1, x_2, \ldots, x_k$ of the independent variables can be expressed as:

$$\text{logit}\left(\hat{p}\right) = \hat{\beta}_0 + \hat{\beta_1} x_{1i} + \ldots + \hat{\beta}_k x_{ki}$$

Or

$$p_i = \frac{\exp\left(\hat{\eta}_i\right)}{1 + \exp\left(\hat{\eta}_i\right)}$$

In epidemiological terms, such probabilities explain the risk of an individual at risk (Carney et al., 2013, Davaalkham et al., 2006). To know about the precision of the estimated β-parameters in a logistic regression, the standard errors of the estimate, *S.E* ($\hat{\beta}$) should also be computed (Collett, 2002). Ghias and Pervaiz,(2009c), Zakizad (2009) and Gheorghe *et al.*, (2010) fit binary logistic regression model on their data sets for the similar purpose and objectives.

3.13.2 Inference for the Logistic Regression Coefficients

On successful completion of estimation of logistic regression parameters, next phase starts with inference of the variables in the model. This usually involves testing of statistical hypothesis for investigating which variables are significantly associated with the response variable. The main purpose of this inference is to know "Does the model deliver more information for the outcome with the variables that included in the model than those which could not be included". This can be achieved by comparing observed values to the predicted values obtained through models with and without the variables. In logistic regression, such sort of comparison can be achieved using the following *likelihood ratio test*: The other method includes Wald's and Score tests respectively.

$$\boldsymbol{D = -2\ln\left[\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model}\right]}$$

Or

$$\boldsymbol{D = -2ln\left(\frac{L_0}{L_1}\right) = -2[ln(L_0) - ln(L_1)] = -2(L_0 - L_1)} \qquad \textbf{3.6}$$

Where D is known as deviance function of the model and presents similar role as sum of square of errors (SSE) manifests in simple linear regression. This likelihood-ratio test also representing as $-2LL$ which helps in determining the parameter values that maximize the likelihood function under the hypothesis $H_0: \beta = 0$, when H_0 is false indicates that the ratio of maximized likelihoods be apt to far below one (Agresti, 2007). The deviance is also used in evaluating goodness of fit of the model but here discussed in the context of assessing the significance of a variable. For the univariate case, the above equation **3.6** can be summarized as:

$$\mathbf{D = -2\sum_{i=1}^{n}\{p_i logit(p_i) + log(1 - p_i)\}}$$

3.7

Equation **3.6** can also be represented as $D = -2\ln(likelihood\ of\ the\ fitted\ model)$ as likelihood of the saturated model is 1 for the binary outcome variable (Collett, 2002).

The "G" statistics determines the change in deviance on the inclusion of the independent variable and obtained as follows:

$$\mathrm{G} = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right]$$

The G statistics follows a χ^2 distribution with v degree of freedom. A p-value < 0.05 indicates that a particular variable is a significant variable and contribute effectively in predicting outcome (Hosmer and Lemeshow, 2000).

Another most widely used test is the *Wald Test*, obtained by taking ratio of the maximum likelihood estimate of the parameter to its standard error. Yazdanpanah *et al.*,(2005) also used the same test "Wald Test" in their case control study from European population for evaluation of regression coefficient inference. Other studies by different authors also imply this test while constructing LR model, for example, Pérez *et al.*, (2005), Ghias and Pervaiz,(2009c) and Southern *et al.*, (2011), A univariate case may be expressed as: $W = \frac{\hat{\beta_j}}{SE(\hat{\beta_j})}$ or $Z = \frac{\hat{\beta_j}}{ASE}$; where, $\hat{\beta_j}$ is the maximum likelihood estimates of all parameters β_j and ASE is the asymptotic standard error of $\hat{\beta}$. The resultant ratio follows the standard normal distribution as well. To achieve the maximum likelihood estimates, complex iterative procedures are required for large number of variables. Both the likelihood ratio test and Wald Test require the calculation of maximum likelihood estimate for the parameter. The *Wald Test* evaluates the significance of each independent variable individually under null hypothesis: $H_0: \beta = 0$ and suggests that whether independents variable are significantly associated with the outcome or not at p-value (5% level of significance) (Hosmer and Lemeshow, 2000). Agresti (2007) recommends likelihood ratio test instead of Wald Test to measure the significance of individual coefficient when sample size is small and for backward elimination. Menard (2002) also pointed out that the value of Wald Test diminishes with large coefficients and their large standard errors. The multivariable equivalence of the *Wald Test* is attained from the following vector-matrix:

$$W = \hat{\beta}'[\hat{Var}(\beta)^{-1}]$$

This will be distributed as χ^2 with $k+1$ degree of freedom under H_0 that all coefficients assume the value zero (Hosmer and Lemeshow, 2000).

3.13.3 Interpretation of Logistic Model Parameters

Importance of applying the logistic regression on epidemiological data was due to an easy interpretation of coefficients in the model. The odds of occurrence of a disease in an individual who is expectedly unexposed are $\frac{p_1}{(1-p_1)} = \exp(\beta_0)$. Likewise, the odds of disease in an exposed person are $\frac{p_2}{(1-p_2)} = \exp(\beta_0 + \beta_1)$ and the ratio of the odds of disease is as follows:

$$\psi = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_2}{(1-p_2)}} = \exp(\beta_1)$$

As a result β_1 is the log odds ratio. This states that in a logistic regression model, regression coefficients give the estimated change in the log-odds on a unit change in the subsequent independent variable assuming that other variables are constant. For example, a logit coefficient of 0.35 reveals that log-odds augment by 0.35 for every one unit increase in the predictor variable (Allison, 1999). The estimated parameters are generally exponentiated to provide results in the form of odds ratio (Landau and Everitt, 2004). Odds ratio (OR), is the parameter of interest in a logistic regression because of its ease of interpretation and numerous software packages are readily available to ascertain its estimates including point as well as confidence interval estimates in a fitted model (Hosmer and Lemeshow, 2000, Collett, 2002).

3.13.4 Model Checking and Diagnostics

The task for model fitting is not so simple and requires great attention of the researcher. Once the model fitted to the data, it is indispensable to check its adequacy through the statistical procedure collectively known as *diagnostic checks*. For this purpose, researchers need to do some additional work in evaluating goodness of fit, checking of multicollinearity, examination of outliers and so forth. This can provide an additional opportunity to the researcher to look into the fitted model once more to improve the deficiency, if any. New model have to fit until a model with good performance not achieved (Rossi, 2009).

a) Measurement of Goodness of Fit

As it has been mentioned earlier that researchers mainly desired to reach on a model which contains good features or predict the outcome in a best possible way. For this purpose, some practical methods for assessing the extent to which a model predicts the observed

outcomes are often known as goodness of fit of the fitted model. This explains that how observed and predicted values are closer to each other (Kleinbaum *et al.*, 2002). In simple words, goodness of fit test is the adequacy of the fitted model and different statistics are available to review model adequacy which are presented below. If this agreement between observed and predicted values is not well established then model should be reiterated accordingly. The null hypothesis to test goodness of fit is defined as H_0: The model fits adequately; alternative hypothesis is H_1: The model does not fits adequately. P-value>0.05 indicates model fits adequately (Kleinbaum *et al.*, 2002). Wise *et al.*, (2010) also find goodness of fit test for LR model in their case control study.

b) The Hosmer-Lemeshow Statistic

Wise *et al.*, (2010) carried out a case control study to evaluate prognostic risk factors of HCV infection in USA and applied Hosmer-Lemeshow Statistic for measuring goodness of the fit test of LR model. Similarly, another similar study from Brazil also implied the same test for testing model adequacy (Oliveira-Filho *et al.*, 2010).This is most widely used statistics for the purpose. To compute the HL statistics, initially, predicted probabilities for all "**n"** individuals are to be calculated. Thereafter, these predicted probabilities are arranged in an ascending or descending order. Then these ordered probabilities are usually divided into deciles of risk groups. Afterward, preparation of a table from observed and expected counts is managed. Finally, HL statistics is evaluated by the formula given in equation **3.8** and compared across the groups with χ^2-statistics at g-1 degree of freedom (Kleinbaum *et al.*, 2002). If the chi-square statistic is not significant then the model fits adequately, otherwise not. Suppose that m_i are the observations in the $i^{th}$of g groups, o_i & $e_i$the observed and expected counts, respectively.

$$X^2_{HL} = \sum_{i=1}^{g} \frac{\left(o_i - m_i\hat{\pi}_i\right)^2}{m_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)} \qquad \textbf{3.8}$$

Also, let $\hat{\pi}_i$ be the average success probability in the i^{th} group, so that $\hat{\pi}_i = e_i / m_i$. The HL-statistic is then given as above (Collett, 2002).

c) Analogues of the R^2-statistics

Other measures of goodness of fit embrace analogues of the R^2-statistics. In linear regression analysis, the most broadly used statistics is the R^2 and known as the *coefficient of determination.* Its value lies between 0 and 1 and expressed in percentage to explain that how much variation in response variable can be explained by the explanatory variables. In logistic regression equivalent statistics like Cox & Snell $\boldsymbol{R^2}$ and Nagelkerke's $\boldsymbol{R^2}$ named as Pseudo$\boldsymbol{R^2}$. Different studies in literature used these statistics for the similar objectives. For example, Ottenbacher *et al.*,(2004), Irfan *et al.*,(2011) and Wang *et al.*,(2013) applied the same statistics while fitting logistic regression model to their data sets.

d) Cox & Snell's R^2_{cs}

Cox and Snell's proposed this statistic in (1989) to calculate the proportion of unexplained variance that is reduced by adding variables to the model.

$$R^2_{CS} = 1 - \{\left(\frac{\hat{L_0}}{L(\hat{\beta})}\right\}^{\frac{2}{n}}$$

Where $L(\hat{\beta})$ and $\hat{L_0}$ are the maximized likelihood of the fitted model and the maximized likelihood for the null model, n is the total number of binary observations. Ottenbacher *et al.*,(2004), Irfan *et al.*,(2011) and Wang *et al.*,(2013) also implied this statistics.

e) Nagelkerke's R^2

The major problem with Cox and Snell $\boldsymbol{R^2}$ statistic is that it remains unable to reach its maximum value of 1, making it hard to interpret. That is why Nagelkerke in (1991)developed a modified form of Cox and Snell $\boldsymbol{R^2}$ that ranges between 0 and 1 similar to the multiple linear correlation coefficient and usually expressed in percentage. The Nagelkerke's $\boldsymbol{R^2}$ is generally higher than Cox's(Peat and Barton, 2008).

$$R^2 = \left\{\frac{R^2_{CS}}{1 - (\hat{L_0}\ )^{\frac{2}{n}}}\right\}$$

In SPSS on running logistic regression model, these statistics are compulsory part of the summary table and measures the strength of the association. This statistics was also applied by Ottenbacher *et al.*,(2004), Irfan *et al.*,(2011) and Wang *et al.*,(2013).

3.13.5 Measuring of Multicollinearity

Before concluding the final model, multicollinearity or collinearity of covariates should also be assessed in the logistic model. This occurs when there are strong linear dependencies prevail among the predictors which can badly affect the regression parameters by raising their estimated coefficients and associated standard errors abnormally. Field (2000) suggests that logistic regression is equally essential to test for collinearity diagnostics. Different criterions were proposed for evaluating collinearity in logistic regression. For instance, Menard (2002) suggests that a tolerance limit value less than 0.1 depicts sever collinearity problem. Myers (1990) also suggests that a Variance inflation factor (VIF>10) is attention-able. Whereas Allison (1999) recommend a threshold of (VIF<2), indicating no significant multicollinearity. Other statistics like *condition index* may also be useful for identification of collinearity in the variables (Field, 2000). Checking of multicollinearity in epidemiological studies was lacking in the literature. However, Nokhodian *et al.*,(2012) considered this issue of multicollinearity in their similar study. In our recent article, this issue was also taken into the consideration (Ghias et al., 2012b).

3.13.6 Residuals and Outliers Analysis

Wise *et al.*, (2010) also find out outliers in a LR model in their case control study. Residuals imply that a measure of agreement between the observed and predicted values, providing much information about the adequacy of the fitted model. The graphical representations of these residuals are highly informative for identifying outliers in the data, explicitly when the data is collecting from observational studies. The outliers are the observations that have shown abnormally large residual and surprisingly distant from the other values (Collett, 2002). Pregibon (1981) was the poineer to introduce the residual analysis in logistic regression to identifying outlying observations by proposing *deviance residual*. He suggested that a maximum likelhihood fit of a logistic regression model is extremly sensitive to outlying observations and such values require special attention of the researchers to model the data in a best scientific manner. Outliers detection increase the accuracy of the fitted model (Nurunnabi and West, 2012).

The simplest graphical representation is a plot of the residuals against the corresponding observation number, or case ID and is referred as *index plot*. While a more informative graph than index plot is a plot between the residuals and the values of the linear predictor (Collett,

2002). In our recent study, these graphs of residual analysis were presented for identification of outliers (Ghias et al., 2012b).

Numerous residual methods are available in literature, however only few useful residuals are described below. It is suggested that researchers should apply these residuals plots to locate the abnormal behavior of the data which occurred due to the presence of outliers. The models may be reiterated several times to reach on best fitted model. In epidemiological studies, little has been discussed with reference to outliers. However, this procedure for identification of outliers was also discussed in our earlier papers, Ghias and Pervaiz (2009c) and Ghias *et al.*, (2012b).

a) Pearson Residuals

Normally, a large difference between y_i and $\hat{y}_i$ depicts less precision and vice versa, this precision can be obtained through the standard error; $S.E(y_i) = \sqrt{\{n_i \hat{P}_i (1 - p_i)\}}$.

$$X_i = \frac{y_i - n_i \hat{p}_i}{\sqrt{\{n_i \hat{p}_i (1 - \hat{p}_i)\}}}$$

These residuals are known as *Pearson residuals* (Collett, 2002). These were also applied in our recent papers, Ghias and Pervaiz (2009c) and Ghias *et al.*, (2012b).

b) Standardized residuals

$$r_{p_i} = \frac{y_i - n_i \hat{p}_i}{\sqrt{\{\hat{V}_i (1 - h_i)\}}}$$

where h_i is the diagonal element of the weighted hat matrix. $\mu\hat{}_i = n_i \hat{p_i}$ and $V_i\hat{} = \mu\hat{}_i(n_i - \mu\hat{}_i)/n_i$. These residuals are simply the *Pearson's residuals*, X_i divided by $\sqrt{(1 - h_i)}$, and are therefore known as *standardized residuals* (Collett, 2002). These residual analysis was applied in our recent articles, Ghias and Pervaiz (2009c) and Ghias *et al.*, (2012b)

c) Deviance Residuals

Another type of residual can be obtained from the deviance, given by

$$d_i = \text{sgn}(y_i - \hat{y}_i)\left\{2y_i \log\left(\frac{yi}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}^{1/2}$$

Where $\text{sgn}(y_i - \hat{y}_i)$ is the function that makes d_i positive when $y_i > \hat{y}_i$ and negative when $y_i < \hat{y}_i$. The quantity d_i is known as a *deviance residual*. This can also be standardized by dividing $\sqrt{(1-h_i)}$ to give *standardized deviance residuals* (Collett, 2002), given as:

$$r_{D_i} = \frac{d_i}{\sqrt{(1-h_i)}}$$

Some other residuals like normalized residuals, logit residual may also be considered additional to the aforementioned residuals. Cooks distance and leverage values can also be computed and compared through the same sort of graphs. These residuals are readily available in SPSS, Statistica, STATA and SAS computer software. These residuals were also applied in our publications, Ghias and Pervaiz (2009c) and Ghias *et al.*, (2012b).

3.14 Neural Network Models

Neural networks or Artificial Neural Networks (ANN) are models that inspired to human brain and have ability to solve complex problems and most widely used method for predicting binary response in medicine. It usefulness from literature is described in section 2.4, however, it is added that ANN models are data mining, non-parametric techniques successfully used in many research areas including classification, prediction, pattern recognition, clustering, forecasting and so forth due to its internal power, flexibility, and ease of use (Samarasinghe, 2006). Tu (1996) described that in a neural network, similar objectives are achieved as in logistic regression modeling i.e. predicting the outcome based on potential predictors. Nevertheless, the model building approach and strategy is dissimilar. Neural networks do not require distributional assumptions or collinearity of input variables like the conventional logistic regression model (Zurada and Lonial, 2011). Maimon and Rokach (2005) explained that Neural networks are closely related to statistical methods and are considered as the complement of statistical techniques, such as discriminant analysis and the logistic regression. An exciting research paper by Tu (1996) emphasized the importance of Artificial neural networks (ANN) as an alternative to logistic regression (LR). He explains that ANN models have ability to detect all possible interactions effects between different exposures with less formal statistical assumptions. However, being the black box and prone to over fitting might be discouraged.

Several types of neural networks that have been proposed in literature, however, the most commonly studied artificial neural network (ANN) model is multilayer feed-forward neural networks, also called *multi-layer perceptrons (MLP)* (Maimon and Rokach, 2005, Lippmann, 1987). This method works in a supervised mode and transfer information only forward direction, from input to output as explained in

Figure 3.2. They also learn how to transform input data into desired response. The network can be expressed as follows:-

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)} f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

Where y_k shows the kth output, f_{outer} output layer transfer function, f_{inner} input layer transfer function, w indicates the weights and biases, (i) the ith layer.

The scaled conjugate gradient method may be used as the optimization technique which trained the network efficiently. The second layer activation is formed when outputs from the hidden layer are connected via weighted connections to the output node and bias.

$$\boldsymbol{a_j^{(2)}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{h}} \boldsymbol{w_j^{(2)}}\, \emptyset_{\boldsymbol{j}} + \boldsymbol{b^{(2)}}$$

This 2nd layer activation is transformed by the logistic of sigmoid output activation function i.e. $y = \frac{1}{1+e^{-a(2)}}$. Later one, the MLP network is then trained for estimation of weights. This training process is usually based on iterative techniques as the non-linear estimation is not straightforward (Leke-Betechuoh et al., 2006).

Some authors have applied ANN model to evaluate risk factors of different diseases and described as a powerful alternative to conventional logistic regression model (Qin et al., 2005, Voss et al., 2002, Gao et al., 2004). These models are appropriate for establishing non-linear complex relationships between a set of input variables and one or more output variables. Sherriff and Ott (2004) carried out a research to identify risk factors for early infant wheeze and used ANN model for this purpose. Another study by Tabaton *et al.*,(2010) also use the same model on different data set and disease. The two methods have been compared in studying risk factors of other diseases; for example, in diabetes, Salmonella typhimurium infections, and coronary artery bypass (Qin et al., 2005, Voss et al., 2002, Gao et al., 2004). These studies showed that ANN model can be a useful alternative method of LR model and applied for the similar objectives. A brief about model architecture and training networks procedure are given in the following subsection.

3.14.1 Model Architecture

Different authors implied Multi-layer Perceptran method for modeling and analyzing risk factors data in ANN modeling technique. For example, Xue *et al*., (1996) applied this technique for risk factors identification. Redondo and Espinosa (1999) and Mohamed *et al*.,(2011) also applied the similar method for modeling ANN model on the risk factors data.

Figure 3.2 reveals the mechanism of multi-layer Perceptran (MLP) model which consists of three layers i.e. Input, hidden and output layers. Input layer includes possible list of independent variables which particularly included for research purpose. The middle layer is called “hidden” and contributes to increase the model flexibility. The third one is the output layer and specified for the target or dependent variable. Each layer has different nodes and

these nodes are interconnected with connection weights, obtained by a series of a mathematical equations (Tu, 1996). These connection weights carry important information through the network. At each hidden and output node, a weighted linear combination of the input variables summed and then a logistic or sigmoid transformation is applied as an activation function. These weights or synaptic weights represent association (and possibly causation) between the observed predictors and outcome (Taylor *et al.*, 2002).

Detail mathematics are avoided, however, a brief illustration of the model is given here. All inputs are multiplied with a corresponding input weight and summed *i.e* $u_i = \sum w_{ij} x_j$. Then this sum is transformed through the logistic activation function, as our response variable is binary. The function of each node is performed as $Out_n = f(\sum_i w_i x_i)$. Where Out_n is the output of a particular node and f is the activation function, even if there are many activation functions are available depending upon the choice of the output variable. The logistic function is the optimal choice for categorical variable, often used for classification problems to limit the range of the network outputs. In this study output variable is binary (cases/controls), therefore, sigmoid or logistic function is a suitable option. Hence, $f(x) = (1 + \exp(-x))^{-1}$. Further, let $X = (x_1, x_2, \ldots., x_k)$ be a vector of k input variables, $Y = (y_1, y_2, \ldots., y_d)$, be the d-dimensional output vector. Moreover,$w_1$and $w_2$be the matrices of connection weights which transfer the important information from the input to hidden layer, then hidden to output layer, respectively. Resultantly, a three layer neural networks can represented as $Y = f_2(w_2 f_1(w_1 X))$ where $f_1$and $f_2$are the logistic functions for hidden and output nodes, respectively.(Maimon and Rokach, 2005)

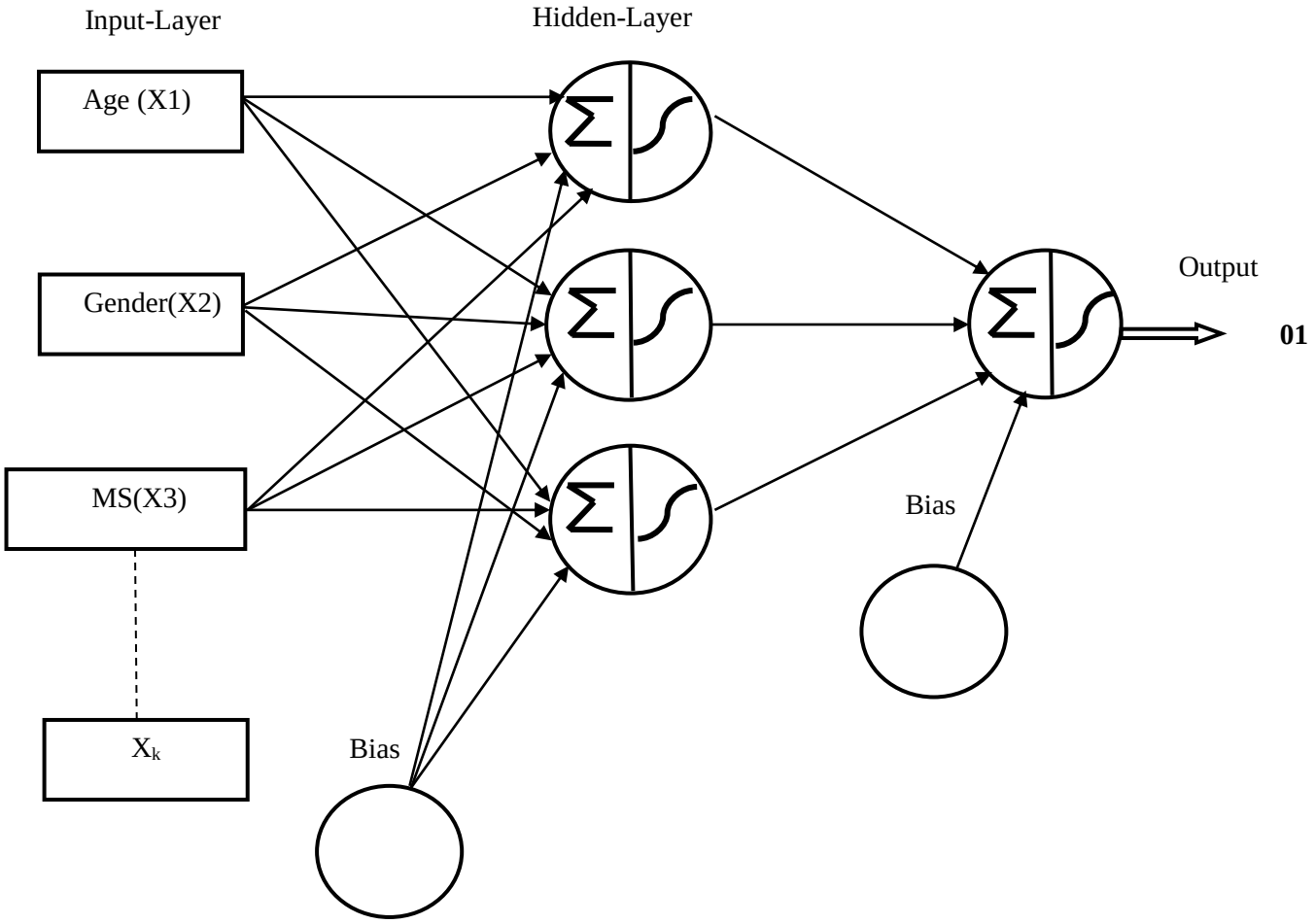


Figure 3.2: Multi-layer Perceptran (MLP) Model Diagram

3.14.2 Networks Training

Networks training demonstrate that how the network should be trained for estimating connection weights (synoptic weights). These weights are identical to the parameters in linear regression. The process is of much importance and entails great consideration of the researchers. The network training is primarily an unconstrained nonlinear optimization problem, performed to obtain the set of weights that minimize the objective function. This process requires some numerical methods to solve nonlinear optimization problems. For this purpose, different optimization algorithms have been proposed to estimate these synoptic weights under appropriate training schemes. The gradient back-propagation algorithm is the oldest one and commonly used method, particularly for large data sets. However, it is

considered a less reliable method due to its sensitivity to local minima (Tufféry, 2011). To compensate the situation, the other method includes *Gradient descent* (Apolloni *et al.*, 2002), *Scaled conjugate gradient* (Møller, 1993) and the *Levenberg–Marquardt algorithm*. Tufféry (2011) suggested that the conjugate gradient descent algorithm is a better option due to its better performance in terms of convergence. Wong *et al.*,(2003) presented a study wherein performances of LR and ANN were combine examined on medical outcome data. The authors also implied same conjugate gradient descent approach for estimation of synoptic weights. Another study by West *et al.*,(1997) also applied the same method and stated that the method Gradient Decent is used for updating the weights and to minimize total aggregate error. Hagan *et al.*, (1996) explained that the use of suitable optimization algorithm facilitates to find proper set of connection weights in such a way that error between predicted and outputs is reduced. Such algorithm requires an iterative procedure which can be implied using batch, line and mini-batch training method. Among these, the batch training is ideally chosen because it directly reduces the total error and updates the synoptic weights after fleeting all training data (Tufféry, 2011). The network starts it's learning by assigning a random value to each of the weights and find out for the known expected value. Then sum of squares of the errors against each single individual is computed by comparing output and the expected values i.e. $\sum_i \sum_j (e_{ij} - o_{ij})^2$

Where, the 1st summation is applied on the subjects of the learning set and 2nd summation on the output nodes. This function enables the network to adjust for the weights after observing tentative increase or decrease in the errors (Tufféry, 2011).

3.14.3 Model Parameter Estimation

In order to parameter estimation, the maximum likelihood estimation is used both for LR and ANN models. In this method, parameters α are selected to maximize the function i.e. $\Pi_{i=1}^{n} P(y_i / x_i, \alpha)$. The neural networks are usually trained by minimizing an error function i.e. *cross-entropy error*

$$\sum_{i=1}^{n} y log o_N + (1-y) \log(1 - o_N)$$

3.14.4 Variable Selection or Risk Factors identification by ANN Model

Logistic regression model has unique importance due to very straightforward variable selection methods, for example, *forward selection, backward selection, and stepwise selection*. However, artificial neural net-works exhibits non linear nature and the statistical tests for parameter significance cannot be applied (Dreiseitl and Ohno-Machado, 2002). Moreover, the ANN model does not permit to interpret coefficients (weights) of each predictor variable or do not carry real-life interpretation like LR model, and hence known as "Black Box". This can be compensated by applying sensitivity analysis, which computes the relative importance of each input parameter in the model and help to rank the variables in order of their importance. The importance of an independent variable measures that how the network's model-predicted value varies for different values of the predictor values. Similarly, the normalized importance is the useful ratio of importance values which are estimated by the ANN model and the largest values, the expressing in percentage. This is a useful way to rank the variables in order of their importance. Variables are ranked in ascending order and represented in bar chart.

3.14.5 Model Evaluation

The commonly used measures of discrimination are sensitivity, specificity, and the area under the ROC curve. These measures apply statistical tests to determine the model discrimination ability (Dreiseitl and Ohno-Machado, 2002). The receive operating characteristic (ROC) methodology is also a useful computational methodology for building and interpreting neural network model. It allows us to assess the overall performance of the model by computing area under the ROC (AUROC). Thus, the sensitivity analysis may help in comparing different models as well selection for pertinent risk factors. Regarding interactions, the ANN model does not show interactions physically in the model, however, interactions are considered by the model which remains hidden. Overall predictive strength of the fitted model may be increased due to inclusion of all complex interactions.

3.15 Classification Trees Model

Keeping in view objectives of this study, Classification Tree models referred as "Decision Trees" methods are emerged as a powerful classification tool due to rapid research in data mining methods (LI *et al.*, 2012). An application of these models through the useful literature is described in section 2.5. However, it is added that these models help the researcher to interpret the results visually in a graphical pattern or diagram as illustrated in

Figure 5.3. Different Decision Tree methods based on specific recursive portioning criteria are available in literature, for example, C4.5, CHAID (Chi-Square Automatic Algorithm for Detecting Interactions), CART (Classification and Regression Tree) and QUEST methods. Among these portioning criteria, the CART model is the famous in modeling and analyzing the risk factors data (Lemon et al., 2003, Hasford et al., 1993, Ture et al., 2005). The authors suggested that CART model can be used as the complementary tool to regression models which identifies pertinent relationship between the predictors and dependent variables at different levels.

An other latest study by Rastegari *et al.*,(2013) also applied CART model to determine risk factors influencing injecting drug abuse in Iran and concluded that CART model gives simplest way of interpretation to the researcher by adding interaction effects. If the dependent variable is categorical in nature it is known as classification trees while with continuous variable referred as Regression Trees. The CART model was invented by Breiman *et al.*,(1984) and help in prediction and classification mainly. It carries only binary partitions or splits of the variable. The most well-known splitting criteria used in CART model is Gini Index, assess the in-homogeneity in binary splits and select the best separation of each node. The best split on a variable will be the one that minimizes the Gini Index (Rastegari *et al.,* 2013). The structure of the CART tree diagram is explained in **Figure 5.3.** There are different types of nodes i.e root node, child node and terminal node which can been seen from this diagram. Root node comprises of total sample size which later on splits into different child nodes by following the splitting criteria. The processes of splitting the nodes remain continue until the minimum termination criterion does not reach. In CART model over fitting may occur if size of the tree becomes enlarge but a too short tree does not represent the true picture of the data. It is therefore, pruning of the tree diagram is recommended. Pruning is the process to shorten the tree diagram which starts from the terminal node and this node will be deleted if its elimination causes a misclassification cost which is significantly lower than the reduction in complexity. This process remains continue until optimal tree is obtained.

Lemon et al.,(2003) has given an impressive methodological review of Classification Trees and LR models in Public Health. They have stated that CART model is not only the promising tool for the identification of at-risk populations but also enable us to produce a visual output in a multilevel structure. They further expressed that the C&RT procedure examines all possible independent variables and chose only those who are most different with

respect to the outcome variable (Lemon et al., 2003). This procedure is named as *splitting criteria* which may help to select the best combination of categories under predetermined splitting criteria.

3.15.1 Splitting Criteria

Different authors have proposed different splitting criterions. Each splitting criteria is based on functions $p_{i/j}$, also known as *impurity function*. There are different impurity functions like Gini, minimum error and entropy. Main function of the splitting criteria is to select "optimal split that has the largest difference between the impurity of the parent node and a weighted average of the impurity of the two child nodes" (Lemon et al., 2003). When the dependent variable is categorical in nature, the most widely used splitting criteria is the Gini. The Gini Index can be calculated in four steps, given as below:-

1. Firstly, the Gini impurity function or diversity index is calculated for the parent node:-

   Diversity index= $1 - \sum p_{i/j} \quad 2 = 2p_{\frac{i}{j}}\left(1 - p_{\frac{1}{j}}\right)$

2. Secondly, calculation of Gini diversity index is made for each two child nodes of the parent node which is under splits.
3. Thirdly, the weighted average of the Gini diversity index is calculated as:-

   Weighted diversity index=$[(p_1)(diversity\ index_1)] + [(p_2)(diversity\ index_2)$

   Where p_1 and $p_2$refer to the proportions of the parent node which are considered in the respective child nodes.
4. The last step, calculates the Gini improvement index:-

   Improvement index=diversity index of parent node-weighted diversity index

The independent variable whose splits possess largest value of improvement index would be selected for splitting at each step.

3.15.2 Stop splitting Rules

After determining the suitable splitting criteria, there is a need of appropriate way of stopping the process. The splitting process remains continue until the entire terminal nodes are homogenous, if there are no criteria. The CART allows the researcher to priori set the criteria for stopping the tree growing procedure, named as *Stopping Rules*. The researchers may specify how large the tree be constructed. The researchers may define minimum

number of individuals in the child/terminal nodes or define maximum number of levels to which the tree can grow. This leads to determine how large the tree should be constructing.

3.15.3 Pruning

Stopping criteria as described above, may not allow growing the tree with full potential. This can be compensated with adopting 'pruning' technique. It allows, initially growing a drastically large tree by including many levels and nodes. Later on, trimming (the 'pruning') of the large tree can be done by ignoring insignificant branches, providing redundant information. Pruning may overcome the o*verfitting* in model building, even in CART model. This phenomenon arises when large number of independent variables might be included in the model, with insignificant effect.

3.15.4 Assessment of Model Fit

Prediction models are generally constructed for prediction of a new entry data. It is therefore, researchers are more interested to build a model which possesses precise, as good as possible, prediction. It is almost impossible to achieve absolute prediction. However, efforts should be made to achieve an optimal prediction level or minimum erroneous classification costs. Simply, a better prediction would entail more reduced rate of wrong classification. Predictive accuracy or assessment of the CART model is made by observing, sensitivity, specificity or area under the ROC (AUROC) curve. The same tools were calculated in each model LR, ANN and CART and performance was compared.

Chapter 4

Descriptive Analysis
&
Application of Logistic Regression Model

The main objectives of this study were to model and analyze the risk factor of hepatitis C in the Punjab province of Pakistan using differential but useful statistical modeling techniques. In this chapter, the data were analyzed descriptively and analytically. At first, a descriptive analysis is performed separately for cases and controls. Later on different multiple logistic regression models are run on different settings of data. And application of ANN and Classification Tree models are discussed in Chapter No.5

4.1 Descriptive Analysis

Mean, median, mode, range, frequencies, proportions, and standard deviation of different variables are computed where appropriate. The descriptive statistics of each of the socio-demographic, medical history, behavioral characteristics and family history related risk factors are discussed as under:-

4.1.1 Socio-demographic Factors

In this section descriptive of socio-demographic factors are explained.

a) Gender

Results from overall data reveal that about 761 (54%) are males and 639 (45.6%) females. Souto *et al* (2012) pooled the data of 9 different epidemiological studies and identify the risk factors of hepatitis C in Brazilian population. The pooled sample indicated that 56.8% are male and 43.3% female. Our data reveals that out of 700 controls; 364 (52%) male and 336 (48%) female. Similarly, among the cases the counts (percentages) of male and female are noted as 397 (57%) and 303 (43.3%) respectively. This indicates that disease is more frequent in males compared to females. Hajiani *et al.,* (2006a) supports our finding in a case control study from Iran and reported that about 63% cases are male while 37% female. In the same context, the results are further supported by Wolff *et al.,* (2008) and Souto *et al.,* (2012). Another descriptive study from Pakistan reveals that frequency of HCV-positivity among the males is higher 353 (78.4%) than 97 (21.6%) female subjects. Ahmed *et al.*,(2012)

also reported higher rates of infection among males. However, some contrary findings are also reported; for example, Lee *et al.,* (2011) and Guadagnino *et al.,* (1997) reported that females are more frequent of this infection as compared to males.

b) Age

The age (in years) is taken as continuous variable and patients of all ages are included in the study without any age preference. Souto *et al.*,(2012) also consider this variable as a continuous in their case control study. Our data also reveal that in the overall sample, minimum and maximum ages are observed as 14 and 80 years, respectively, (rage 66 years; mean±SD (34.14± 10.43)). el-Sadawy *et al.*,.(2004) observed a concordant mean age in Egyptian population (34.7%). Moreover, the median and mode of the ages are 34 and 40 years, respectively. A similar range of ages is found in cases and controls with (mean±SD) 35.0± 10.13; median 35 years and mean±SD (33.0± 10.61), median 31 years. The difference between the mean ages of cases and controls was statistically significant ($p<0.01$). While a case control study from Quetta, Pakistan revealed that no significant difference is observed in ages of cases and controls (Ghaffar *et al.*, 2009). The mean ages in males and female are recorded as 33.76 (±10.46) and 34.60 (±10.37), respectively indicating no significance difference ($p=0.132$). Souto *et al.,* (2012) found that ages lie between 2–86, median 34 years.

In order to get the significant age group, the continuous variable is further categorized into two age groups *i.e* ≤34 and >34 taking cut point at mean value. The results indicate that about 53% of cases belong to >34 age group and in addition about 10% cases exposed against age equivalent to 40 years. This depicts that majority of the patients belong to elderly age group. Idrees *et al.*,(2008) and Lee *et al.,* (2011) described that the risk of hepatitis C disease increased sharply with age due to increase in risk of certain exposures, so being older age is an important risk factor. Idrees *et al.,* (2008) also divided the patients ages in two groups *i.e* <35 years and ≥35 years and majority of the cases fall into the older group.

Different authors have compared patients' age groups on different cut point values depending upon their data behavior. For instance, Shaikh *et al.*, (2009) divided the age group into three groups *i.e* <20 years, 20-40 years and >40 years and found that about 58.2% patients belong to age group 20-40 years. Similarly, Qazi *et al*(2011) divided the patients ages in four groups ≤29, 30-39, 40-49 and >50 wherein about 50% cases belong to age group 30-49 years. Bari *et al.*,.(2001) also divided the ages into 4 groups' *i.e* (20-24, 25-34, 35-44 and 45-70) and expressed that majority (45.6%) of the cases stood for the last age

group. A perception may arises that consideration of age group 14-18 may dilute the findings of the study. Upon illustration, this would be evident that the patients of all ages and gender be studied as per pre-defined inclusion criteria. There are about 50 patients under this age group and if removed, sample size would be reduced and study findings may not represent the target population in real prudence. Hence, the exclusion of this age group while analyzing the data may ignore the important age group which still exists in the population.

c) Marital Status

Information regarding marital status is also studied under two categories "Ever-married" and "Never-married" to explore any association with hepatitis C infection. Kao *et al.*,(1992), Akahane *et al*.,(1994) and Habib *et al*.,(2001) explained that marriage is also a strong indicator for HCV infection indicating transmission of HCV in spouse, although chances are bit low. The risk is proportional to the duration of marriage as well. From the overall data, the results suggested that about 296 (21.1%) patients are never-married while 1104 (78.9%) ever-married. The ratio of cases is more in married patients. In contrast 601 (78.9%) cases belong to ever-married category. A similar proportion (78%) of cases having married status is observed in Egypt population by Kandeel *et al* (2012). This suggests that majority of the patients who ever-married are considered at higher risk than others. The results are also supported by Mohamed *et al.*,(1996) who reported that married individuals had a significant higher HCV-seropositivity than non-married. Bari *et al.*,.(2001) carried a case control study among the male adults in Rawalpindi-Islamabad, Pakistan and also studied marital status of the patients. They found that about 87.7% cases are "ever-married" and only 12.5% "never-married" which menifests compatible findings with the recent study.

d) Family Status

Family status includes nuclear and joint family is also studied in this study. In accordance with a Gallup survey in Pakistan, more than two third of all Pakistanis (67%) argue to favor residing in a joint family nevertheless 31% prefer nuclear family. Recently, the joint family system has continued to develop, and there is a greater presence of nuclear families in central Punjab (Lago, 2011). It is suspected that individuals living in joint families are probably exposed to high-risk of hepatitis C as compared to nuclear families. The key reason is that people residing in joint families usually shared belongings of others, for instance, knives, nail cutters, tooth brushes, scissor, clippers, knitting needles etc. If someone in the family has got infection, he/she may transmit this disease to some other person who is

habitually or accidentally shared his/her equipment. Consequently the likelihood or simply risk of hepatitis C may increase substantially in joint families as compared to nuclear.

In this study, 666 (47.6%) and 734 (52.4%) subjects belong to nuclear as well as joint family in the overall data set, respectively. The result depicts that about 385 (55%) cases belong to joint family; conversely, this proportion is moderately low in nuclear family (47.6%). As a result, this variable may facilitate to establish a relationship of family status with the disease in analytical analysis.

e) Residential Locality

This variable is incorporated in the study to evaluate as to how the cases and controls are distributed within urban and rural settings, thus to establish a relationship with the outcome. Further, this variable will certainly help to compare and analyze the risk factors of hepatitis C infection among the patients living in urban/rural settings independently. It is actually believed that certain risk factors predominant in urban/rural settings of patients may differ owing to the differences in medical facilities offered, lifestyle, educational level, and general awareness among the masses regarding spread of disease looks fairly variant.

From Table 4.1, results reveal that out of total sample, 552 (39.4%) individuals belong to urban community while 848 (60.6%) belong to rural dwellings. Similarly, among the cases 404 (57.7%) and 296 (42.3%) patients come out in rural and urban locality, respectively. This explains that majority of the cases belong to rural area and similar is the case with the control group. Mohamed *et al.*,(1996) explains that rural population is more likely to HCV-seropositivity than urban patients. Thus the separate study of risk factors in both dwellings will help in further exploring the pertinent risk factors associated with hepatitis C in Punjab, Pakistan. Ghaffar *et al.*,(2009) carried a case control study to identify risk factors of hepatitis C infection among the women of reproductive age group, in Quetta, Pakistan. They found that About 45.6% cases are from rural communities and 37.9% of the controls are from rural areas. Idrees *et al*(2008) identified that patients residing in rural areas are at high risk and "residence in rural area" is a significant risk factor of hepatitis C infection.

f) House Status

The information pertaining to house status of patients is also gathered in such a way that people are living in their own (owner) or rented houses. This factor may also assist to identify the socio-economic status of the patients. By and large, those people who are residing in their own houses enjoy healthier life style and better socio-economic status as

compared to others having rented houses. In this study, the overall data reveals that about 1101 (78.6%) patients have their own houses whereas 299 (21.4%) residing in rented houses. The data further explains that among the cases, 541 (77.7%) patients have their own houses while 159 (22.7%) living in rented houses. This signifies that majority of the patients have their own houses indicating no apparent association of the disease with individuals having rented houses. This relationship is examined in the forthcoming analytical section and no significant association of "house status" was found with the outcome ($p>0.05$).

g) Family Size

The information regarding "family size" is inquired from the patients that how many persons are living in a house to ascertain average family size in the patients and its relationship with the hepatitis C infection. This variable is treated as a continuous variable and (mean ± SD) is calculated for controls (7.38±3.71), cases (7.53±3.84) and the overall data (7.49±3.76), indicating that average family size is composed of 7 persons in each group. No statistical difference is observed for average family size ($p>0.05$) between cases and controls.

h) Number of Persons Sharing the Room

This variable is included as continuous variable to examine what is the average number of patients who share their room. The descriptive statistics demonstrated that among the cases, mean number of persons sharing one room are 5 persons (±3.08 SD) and for controls this **Figure** wa 3 persons (±1.65), showing a significant difference ($p<0.05$). Abbas *et al*. (2008) introduced a case control study from rural Sindh, Pakistan and analyzed that variable into two categories *i.e* <3 persons and ≥3 persons. They determined that around 32.5% of cases shared the room with at least 3 persons. They added that as the number of persons sharing the room restricted to less than 3 persons category, the risk of disease also remained low. During the data collection, the researcher experienced that in many cases a whole family lived in a single room. Since highlighted above that around 52.4% patients are living in joint families with limited facilities or space in a household. These people are forcefully considered necessary to share the room with various individuals. This reflects a poor socio-economic status in most of the patients.

i) Patient's Education

Undoubtedly, a better education and learning play a pivotal role in controlling and avoiding certain challenges of human being including health concerns. Consequently, it enhances the comprehension of the individuals to develop good sense of awareness for the

known risk factors of distinct ailments. They can independently learn information through literature, reading books or through the internet. Whereas an un-educated person is always reliant on others to get sufficient information and awareness regarding the spread of a particular disease.

In this study patients' education is taken as continuous variable in such a way that "0" indicating no formal education, means illiterate. It is further categorized into 4 different groups' with respect to their years of education *i.e* No Schooling, 1-5, 6-10, 11-16 years to distinguish significant group. It is noticed that majority of the cases (62.7%) are illiterate and had no formal education. In the overall data, the proportion of illiterate patients is observed as 46.8%. The counts (percentages) in other groups are also compared and explicated in Table 4.1. Generally, this represents that literacy level in cases is low as compare to controls. Bari *et al*.,(2001) conducted a case control study from the male adults, Pakistan and categorized patient education by 4 different groups (0, 1-8, 9-12, and >12). They found that more or less 8.8% cases are related to zero education and almost matching percentage (8.3%) is seen in control group. The Islamabad (capital city) of Pakistan had highest literacy rate in Pakistan, therefore, demonstrated low proportion of un-educated patients associated with hepatitis C. In contrast, another study by Ghaffar *et al*.(2009) from the female patients only, belonging to a distant area of Baluchistan province of Pakistan showed that 51.5% female cases had no school education. Shazi and Abbas (2006) revealed in a case control study from Karachi, Pakistan in which risk factors of hepatitis B& C are compared; found that "patient low education" is an important and considerable risk factor. In this study, patients of similar educational background are seen across all divisions of Punjab except Rawalpindi and Lahore, where patients are relatively well educated and responsive. Thus, an education provides a full opportunity to live healthy and successful life by knowing potential risk factors of hepatitis C disease.

j) Monthly Household Income

To compare the frequencies (percentages) of monthly household income in cases and controls, the income groups are divide into 4 groups (in Rs.) as $\leq$ 5000, 5000-15000, 15000-25000, >25000. It is found that about 328 (23.4%) patients in overall sample belong to lowest income group. Similarly, 64.7%, 7.1% and 7.6% patients related to 2^{nd}, 3^{rd} and 4^{th} income group respectively. From the data, it is further observed that almost matching proportion of patients belong to each group. Besides these around 62.3% and 67.1% subjects integrated in 2^{nd} income category. Moreover, mean ± (SD) of monthly household income is

determined in the overall sample, cases and controls separately as; 10278±6969, 9923±6778, 10635±7142, respectively. On the other hand, this indicated that average monthly household income is 104 USD, 108USD and 101USD in overall sample, cases and controls, respectively. A study by Abbas *et al*(2008) from Pakistan showed that monthly household income of hepatitis C patients is 87 USD. They further explained that about 52.8% cases belong to <5000 (Rs.)/per month while present study deems this figure to be 26%. These finding suggest that majority of the patients whether belong to cases or controls, equally hold the poor families. Much of the money they earn is spent on their own food and certainly nothing ended up in saving for health and education. Even though, government hospitals, at some extent, demand bit payment in favor of laboratory tests, diagnostics and surgical operations. Nevertheless, patients professed that they even fail to bear travelling charges to reach the hospital with meager sources. The same situation is observed in majority of the cases in either division of the province, Punjab.

k) Father Education

Regarding father's education of the respective patient, the overall results showed that around 1049 (74.9%) and 351 (25.1%) fathers had and had not formal education, respectively. Among the cases, 81.1% patients answered that their fathers had no education, contrary to that only 25.1% cases reported for educated fathers. These percentages bit different 71.1% and 28% respectively, in our old study (Ghias and Pervaiz, 2009c). It is believed that being the head of the household, an educated father can keep the family updated against the hazardous diseases which may reduce the risk of the disease apparently.

l) Mother Education

Like the father education, mother education of the patients is also very important. It is well recognized that an educated mother can give better health protection and awareness to her family against the fatal diseases, especially to the children. Because she is every time close with her family members and ready to keep them aware in a best possible way subject to her education and awareness. The counts (percentages) showed that about 1134(81%) and 266 (19%) patients reported un-educated as well as educated mother. Concordant findings are reported in our an earlier study (Ghias and Pervaiz, 2009c).

In a nut shell, descriptive study of the socio-demographic factors describes that hepatitis C is somewhat more frequent among the males and older age group. The mean age is 34 years, ranging between 14-80 years. Majority of the cases are usually married and also occupied countryside regions. The trend of joint families is customarily preferred in Pakistan

wherein multiple people have to share the room. This is inevitably susceptible to the disease on sharing personal belongings of each other. Aside from that, the majority of the cases are identified under-educated and even had poor socio-economic status. Fathers and mothers of the respective patients reported to be generally under-educated. Thus, overall descriptive analysis depicts that majority of the patients belong to poor socio-economic status with low level of education.

Akhtar *et al*. (2004) explained that family members infected with HCV and designated low socio-economic status are associated with HCV infection risk. The association of each socio-demographic factor with the disease would be established on statistical grounds in a forthcoming analytical analysis. At present, descriptive of other risk factors including behavioral characteristics, medical and family histories are primarily discussed below.

4.1.2 Medical History Related Factors

In this section risk factor related to patients' medical treatment history and family history are explained.

a) Patient History of Jaundice

Jaundice creates yellow color of the skin and develops with the increase of Bilirubin in the blood. It can be a cause of many different diseases, for instance, jaundice may be a symptom of hepatitis. Blackard *et al* (2007) stated that average incubation period of HCV ranged between 6-7 weeks. He explained that only 10-15% of cases are acutely jaundiced. In this study counts (percentages) of patients in overall sample showed that around 1149(82.1%) patients had no past history of jaundice while only 251(17.9%) reported its positive history. Among the cases, about 150 (21.4%) patients are observed with positive history of jaundice while in a Turkish case control study this percentage is comparatively low (4.6%) (Karaca *et al.*, 2006).

b) Family History of Jaundice

Family history of jaundice is also a considerable risk factor of hepatitis C infection. Ghaffar *et al*. (2009) reported that around 46.8% cases had positive family history of jaundice and lived with jaundiced patients in a household. They claimed that this risk factor had a positive association with hepatitis C disease among the females. The present study established a contradictory percentage (5.6%) of patients with positive family history of jaundice and about 94.4% patients indicated negative history as well. This percentage

comparison is not elucidating any positive association of family history of jaundice with the disease. A similar case control study is conducted from blood donors in France and reported a higher percentage (21%) of cases who had stated positive family history of jaundice (Serfaty *et al.*, 1993).

c) History of Blood Transfusion

At start, it is believed that blood transfusion is the sole contributing factor in the spread of hepatitis C infection. However, recently, many epidemiological studies have proven the importance of other potential risk factors of hepatitis C as well (Ben Alaya Bouafif *et al.*, 2007, Kandeel *et al.*, 2012). Still, it is a very imperative risk factor of hepatitis C infection and most frequently reported in national as well international literature (Karaca et al., 2006, Jafri et al., 2006, Ali et al., 2009a). Karaca *et al*. (2006) reported that about 39.7% patients had history of blood transfusion in Turkish population. The authors further added that about 25% of HCV cases are related to transfusion of contaminated blood and blood products.

In present study, the overall data sample illustrates that around 485 (34.6%) patients have history of blood transfusion. Among the cases, this proportion raises to 44.7% which is match-able with Turkish study by Karaca *et al*. (2006). A study from Germany showed 34.4% of the cases exhibit history of blood transfusion whereas from Rawalpindi, Pakistan, Bari *et al*.(2001) showed comparatively low 16% percentage of blood transfusion. Qureshi *et al*. (2009) and Abbas *et al*.(2008) reported very low proportion (6.3%) of history of blood transfusion but significantly different from controls. The authors stated that the risk of hepatitis C increases 2.9 times with positive history of blood transfusion. This suggests that rate of positive history of blood transfusion decreases with time but the figure (34.6%) of recent study seems to be alarming. However, every author implements the significance of this risk factor and emphasized for its further study.

It has been reported that developed countries have effectively controlled HCV infection in general population by introducing well established blood screening methodology (Alter, 2003, Akhtar *et al.*, 2004). However, in developing countries, the risk of post-transfusion HCV infection is still a dilemma and demands better controlling strategies in future.

d) History of Blood Donation

The counts (percentages) for history of blood donation are displayed both in cases and controls. These mentioned that 1104 (78.9%) never reflected any positive history but also 296

(21.1%) deemed to be positive. Moreover, among the cases 25% patients informed they have already donated blood to a couple of patients just for saving their lives. The males are reported for blood donation solely but no female affirmed blood donation. On further inquiry, the researcher investigated that most of the patients donated blood for the family members, relatives and friends. Generally, the screening tests for HCV and other infectious diseases are done but nevertheless it is far from the mark, specifically in the rural hospitals. Akhtar *et al*.(2004) reported a case control study to determine hepatitis C infection among the blood donors in Karachi, Pakistan. The authors described that the overall seroprevalence of HCV in blood donors is 1.8% and this proportion had been increasing significantly. They further stated that about 43.2% blood donors had donated blood more than 2 times. An analogous study by Liu *et al*.(2009) 10.1% of HCV positive patients have history of blood donation. Thus present study reflects comparatively greater percentage of blood donations than neighboring country China.

e) Dental Surgery

History of having undergone dental surgery is other possible risk factor for infection and is reported by some authors in different regions (Kim *et al.*, 1996, Lasher *et al.*, 2005). The present study finds that about 321 (45.9%) of cases had history of dental surgery. Conversely, controls have relatively low proportion of positive history of dental surgery. Majority of the controls (75.9%) do not indicated any history of dental surgery. However, on comparing these percentages with the cases and controls, a significant difference is identified. During data collection, the researcher noticed that most of the patients have dental surgery from the government hospitals which depicted a frightening situation. It pointed out an improper sterilization of equipments in the general public hospitals in the province Punjab. Idrees *et al*.(2008) found that around 20.3% of cases have history of dental sugary in province, Punjab which is comparatively low as reported in present study. Another study from district Hafizabad in the Punjab by Luby *et al*. (1997) mentioned this proportion (33%) among the cases. Similarly, Qureshi *et al*.(2009) found that history of dental filling, extraction and scaling is significantly higher (38%) in male cases than controls (20.9%).

f) Kidney Hemodialysis

The counts (percentages) of history of kidney dialysis both in cases and controls revealed that none of the control had indicated its positive history. However, only one patient among the cases is reported with history of kidney dialysis. The main reason is that generally

the patients of hemodialysis visit an independent center, reserved for hemodialysis patients only within the hospitals. Therefore, the patients with history of hemodialysis could not be found in outpatients, inpatients or hepatitis clinic from where this study sample is collected. But still, it is a considerable and recognized risk factor of hepatitis C infection which should be studied in a separate study from the hemodialysis patients. In a case control study, Hajiani *et al*.(2006a) found that hemodialysis is an idependent risk factor of HCV infection in Iran. In contrast, Gheorghe (2010) reported this factor is insignificant in Romania population. Generally, the patients who are involved in continuous heamodialysis found to be associated with increasing risk of HCV infection.

g) History of Accidental Needle Stick

An accidental needle stick injury is another suspected risk factor indicating about 330 (47.1%) of cases with its positive history which is higher as compared to controls (27.6%). Another study by Idrees *et al*. (2008) from the province, Punjab identified this figure as 32% indicating a positive association with the hepatitis C infection. It is observed, from the sampled data, that accidental needle stick injury is more commonly reported among paramedics, tailors, and women who tailored at home, prepared handicrafts and share needles that are contaminated during the course of their work.

h) History of Therapeutic Injections

This is the most common risk factor of hepatitis C infection in Pakistani Population. Pakistan is one of the leading countries where un-necessary therapeutic injections are on top and an average of 13.6 injections are received every year per person (Janjua et al., 2006b). In addition to injection overuse, most injections are provided with previously contaminated syringes, particularly in rural areas by quacks (Janjua *et al.*, 2005). The counts (percentages) of this study reveal that history of injections has a relatively high frequency among both cases (79.9%) and controls (60%). As shown in the present study and also as reported in previous studies, participants who received more injections are more likely to be infected with HCV (Idrees *et al.*, 2008). Comandini *et al*., (1998) reported that history of injections is a significant risk factor of HCV infection. A more detailed discussion would be made in the subsequent model building section after performing analytical statistical analysis.

i) Major/Minor Surgery

This is also an evident and pertinent risk factor of hepatitis C infection not only in Pakistan but also in other developing countries. This suggests poor infection control

standards in health care settings. This study data showed that around 975 (69.6%) patients are not consistent with positive history of major/minor surgery in past while 425 (30.4%) patients claimed its positive history, in the overall data sample. Moreover, among the cases about 40% patients reported history consistent with surgical procedures while in controls this figure reduced to 20.7%. The descriptive results indicate a positive association between surgery and HCV infection. The descriptive statistics of a similar case control study by Bari *et al*.(2001) from male adults in Rawalpindi/Islamabad, Pakistan patients showed a matching percentage (42.1%) of cases who had positive history of surgery. A descriptive study by Hajiani *et al*.(2006a) from Istanbul Turkey reveals that history of surgery is found in 98% patients who had visited gastroenterology clinic for hepatitis C diagnosis/treatment. Idrees and Riazuddin (2008) found that in Pakistan about 70% of the cases are obtained through reuse of syringes and general surgery.

j) History of Angiography/Angioplasty

Ahmad *et al*. (2008) noted an exponential increase in coronary angiography procedures from 2000 -2006 in Pakistan which in turn inspired the researcher to explore its association with disease as previous studies ignored this factor. On this study, it has been hypothesized for the possible infection, when properly sterilized equipments are not implied to perform the surgery. Thus the counts (percentages) of one more suspected risk factor *i.e* "history of angiography/angioplasty" is also considered and results found that around 33 (2.4%) cases have its positive history. The comparable percentages are identified both in cases and controls suggesting certainly no significant association with the disease.

k) History of Endoscope/Gastro scope

Another risk factor in the present study is history of undergoing endoscopy or gastroscope. This association concurs with another study, also from Pakistan, which found that endoscopy is a significant risk factor among women (Hashmi *et al.*, 2010). The present study reveals that a round 107 (7.6%) patients reported a positive history of endoscopy or gastroscope in the overall sample. However, among the cases, the frequency (percentage) of positive history is 84 (12%) which seemed to be significantly different than controls (2.3%). A case-control study, carried out in France (n=1207 including 450 cases) reported that a history of undergoing endoscopy had a significant association with HCV infection, with an estimated odds ratio of 1.9 (Karmochkine et al., 2006b).

l) History of Hospitalization

Kandeel *et al*.(2012) identified that history of hospitalization is the significant risk factor of hepatitis C infection in Egypt. In this study, it is also investigated from every case and control patients that they had ever admitted in hospital or not. The counts (percentages) reveal that around 379 (54.1%) of cases mentioned a positive history of hospitalization which is relatively high as compared to controls 255 (36.4%). The concordant findings are reported by (He *et al.*, 2011) in blood donors population of China. Akhtar *et al*.,(2004) suggests that 58.9% cases do not mention any history of hospitalization while in present study this is going to report as 45.9%. Moreover, they found that among the cases about 41.1% cases depicted positive history of hospitalization which is comparable with findings of present study (36.4%). Qureshi *et al*.(2009) found a contradictory percentage (14.7%) of cases associated with history of hospitalization from the male patients only in Karachi. Akhtar *et al*.(2004) found that cases are 2 times more likely than controls that have past history of hospitalization.

m) History of Branula Insertion

Foster *et al*. (2010) explained that the contaminated IV branula insertion might be the possible cause of infection among the health care's, in case branula needle pricked accidentally. Similarly, this factor has been considered as the suspected risk factor of HCV infection in this study because almost every admitted patient in the hospital ought to experience branula insertion for IV injections/drips etc. The frequency and percentage of history of branula insertion is first time being reported regarding HCV infection. It is observed that about 414 (59.1%) cases are consistent with history of branula insertion, conversely, controls depicts this percentage as (36.6%), suggesting a positive association between the history of branula insertion and HCV infection. In the overall sample, almost 50% patients are identified in either group with positive history of branula insertion.

n) History of Acupuncture

A case control study by Balasekaran *et al*.(1999) is carried out from southern United States to determine the significant risk factors of hepatitis C. The authors also considered the history of acupuncture in their study but results are not significant. Another study from Turkish population indicated that history of acupuncture is infrequent, 2 (0.6%). In Japan, history of acupuncture has no increasing risk of hepatitis C (Nakashima *et al.*, 1993).

In present study, the overall data reveals that only 6 (0.3%) patients have positive history of acupuncture whereas among the cases this figure reduces to 2 (0.3%). Thus, the

present study depicts the matching results with the Turkish study (Karaca *et al.*, 2006). In contrast, a case control study from China indentified that the history of acupuncture is more frequent (16.1%) among cases than controls (3.3%) indicating a significant difference.

o) History of Eye Surgery

Jatoi *et al*.(2009) conducted a study from Hyderabad, Pakistan for evaluating incidence of hepatitis C infection among the patients who underwent eye surgery. The results suggested that about 29.60% patients are HCV positive and most of them are males with male to female ratio of 1.21:1. To our best knowledge, in present study, this particular variable is first time studied as an independent risk factor of HCV and results reveal that about 65 (4.6%) patients affirmed positive history of eye surgery while majority (95.4%) of the patients are negative. However, among the cases the proportion (6%) of patients with positive history of eye surgery is high as compared to controls (3.3%). This describes its relationship with the disease that would be verified in the subsequent section.

p) Ear Wax Removal from Hospital

This particular variable appeared to be first-time in this study as being a suspected risk factor of hepatitis C infection in Pakistani population. It is noticed that in the government hospitals in Punjab province, medical practitioners frequently have to share exactly the same instruments to examine, treat and removing wax in the ears for the patients. Several patients visit these hospitals in a day and proper sterilization is probably not ascertained good enough.

In this study, overall only 29 (2.1%) patients are reported with history of ear wax removal from government hospitals whereas majority of patients (97.9%) mentioned no previous history. Moreover, among the cases around 21 (3%) patients reported its positive history which is bit different in controls 8 (1.1%).

q) Cesarean Section (Female only)

Habib *et al* (2001) in the Nile Delta region, Egypt that Cesarean Section is strongly associated risk factor in female patients. The present data indicated that about 135 (45%) of female cases transmitted HCV infection through Cesarean Section. The trend of cesarean section is going up day by day in Pakistan and the situation becomes more worsen in rural areas where un-trained, un-qualified LHVs/Midwives are performing this procedure without knowing the concept of sterilization. Majority of the female cases received Cesarean Section from the Government Hospital (23%). This factor was found significant both in univariate

and multivariate models. Another study from Pakistan by Ghaffar *et al* (2009) showed insignification association of this factor with the disease.

r) History of Abortion/D&C

This is another important risk factor of hepatitis C prevailing in female patients only. The present data reveal that about 113 (37%) female patients. In univariate analysis, this factor showed its significant association with the outcome and reflected that risk of disease increases by 5 times in females those have underwent history of abortion or D&C. Most of the females could not differentiate between abortion and D & C during data collection, it is therefore, these are taken as combine. Abortion is prohibited in Islam, therefore not so frequent in Pakistan. However, actual data is very hard to find. Majority of the females have to face D & C on different occasions. In rural areas, particularly D & C is also performed by un-trained, un-qualified LHVs/Midwives, therefore, chances of transmission of infection this way augmented. A study from Pakistan by Ghaffar *et al* (2009) reveals an insignificnt association while the recent study possessed positive association. Our data indicates that about 113 (37%) female cases have history of abortion/D & C.

4.1.3 Behavioral Characteristics

a) Tattooing

Habit of tattooing is generally observed in western nations but few studies from Pakistan's perspective experienced its significane with the hepatitis C infection. Balasekaran *et al*.(1999) introduced a case control study from Southern United States and found that tattooing is perfomed often by friends, family members applying unhygienic instruments. They pointed out that about 43% of cases had habit of tattooing. Similarly, He *et al*.(2011) performed a study from China and expressed that tattooing is considerably more frequent (30.5%) in cases than controls (5.6%). A study from Islamabad, Pakistan by Bari *et al*.(2001) exposed that roughly 8.9% of cases practiced tattooing whereas another identical study by Akhtar *et al*.(2004) from Karachi reflects this figure as 7.4%. They documented that in a control group the history of positive tattooing is 5.1% while our study reported as 8.3% which is relatively higher but not strange. Similarly, in present study, the counts (percentages) of positive history of tattooing among the cases is 128 (18.3%) which is significantly lower than United States and China studies. This study sample highlighted that generally males had tattooing practice.

b) Body Piercing

This can be another key risk factor of hepatitis C infection, particularly, women's are profound inspired to display costume jewelry in pierced ears and noses. Some times males may also be engaged in such practices, however, not so common. The women ear/nose piercing is generally performed at local shops in group settings and same contaminated needle is utilized on multiple clients. Akhtar *et al*.(2004) found that body piercing is a significant risk factor of hepatitis C among blood donors of Karachi city, Pakistan. In additon, Idrees *et al*.(2008) reported from provicne Punjab that nearly 9% females had probable mode of transmission by sharing of piercing needle and instruments. In a study from China, He *et al*.(2011) reported that 32.1% of cases had body piercing while in a control group this firgure significantly reduces to 10.2%. From Bangaladesh, Ashraf *et al*.(2010) found that the risk of having HCV infection icreased by 5 times in females having past history of ear/nose piercing. Our data described that 46.9% of cases manifested postivie history of body piercing. Compared to present study, a low proportion (23.2%) of history of body piercing is reported in Korean population (Kim *et al.,* 1996) and China (32.2%) (He *et al.,* 2011).

c) Multiple Marital relation

Kaldor *et al*.,(1992) suggested that multiple partners or more than one marriage is another significant risk factor of HCV. Kao *et al*.,(1992), Akahane *et al*.,(1994), and Habib *et al*.,(2001) further explained that marriage is also a strong risk factor for HCV implicit for transmission through relations of HCV in spouses, although chances are bit low. The counts (percentages) of this expected risk factor reveal that among the cases, only 49 (7%) patients have more than one marriage or multiple marital relation. This figure is relatively low in controls (5.4%) indicating no significant difference.

d) Injected Drugs, Even Once

Injected drug use is another well recognized risk factor of hepatitis C both nationally internationally. Neal *et al*. (1994) performed a case control study out of blood donors of UK inhabitants to analyze risk factors of HCV. They found that about 53% of cases reported history of injected drug which is considerably more frequent than controls. Similarly, another study recorded nearly 45.8% of street young gays had injected drugs. From Pakistan, Idrees *et al*.(2008) stated that injected drug use could be a significant risk factor of HCV in province Punjab. The data of current study reveal that just a small number of cases (7.2%) have

informed positive history of injecting drugs, even once while this percentage is much lower in control group (2.7%). The proportion of injecting drug users is substantially low compared to aforementioned UK based case control study. A recent study in Baluchistan, Pakistan, Ahmed *et al*. (2012) revealed an association between injected drug users and HCV and the risk of HCV increase by 29.95 (95%CI: 7.06–127.02).

e) Reuse of Syringes

A latest case control study carried out by Kandeel *et al*.(2012) reported that reuse of syringes is an ongoing risk factor of HCV in Egypt population. The authors came across that upon sharing used syringes, the unfortunate risk of HCV raised up to 23 times. Idrees and Riazuddin (2008) further concluded that, In Pakistan about 70% of the cases are received through reuse of syringes and general surgery. The counts (percentages) out of present study describes that just 39 (5.6%) of cases received through used syringes. These contradictory details represents to those people who are intended to this act at home or local pharmacy, otherwise, percentage of sharing of used syringes in Pakistan is incredibly high but stayed secret due to criminal act of quacks in rural areas who possibly use already used syringes to save money. This practice is appeared to be more common within the individuals of DG Khan and Sahiwal region of province Punjab.

f) Sharing of Toothbrush/Miswak

The number (percentages) of subjects having habits of sharing of toothbrush/miswak are associated with minor risk of hepatitis C infection. Kandeel *et al*.(2012) reported from Egypt that sharing of toothbrush is common in 9% of cases. In costrast, the present data showed that 26.3% of cases and 18% of the controls are involved in that behaviour in the region.

g) Sharing of Nail Cutter

From Pakistan, Idrees *et al*. (2008), Bari *et al*.(2001) and Akbar *et al*. (2009) reported that sharing of nail clipper is another minor risk factor of HCV. The percentage data of present study indicated a high frequency of sharing of nail cutter is found among the households whether cases 1007 (71.9%) or controls (64.4%). However, the associated risk is considerably low whereas demanding a careful behaviour to circumvent the infection.

h) History of Cuts

History of accidental cuts (from a blade or knife) is another attributable risk factor of hepatitis C (Ghias and Pervaiz, 2009c). The results reveal that about 396 (56.6%) cases have

the history of cuts by the communal unsterilized blade or a knife. In univariate analysis this factor was emerged as a significant risk factor as p-value<0.05. In gender-wise analysis, it is come to realize that history of cuts is more frequent in female (60%) patients than males (53%). This mostly happens when females use the same knife in a kitchen and share it on different times.

i) History of Imprisonment

Delage *et al* (1999), Zakizad (2009) and Gheorghe *et al* (2010) reported that history of imprisonment is a consistent risk factor with hepatitis C infection among the Canadian, China, Romania inhabitants, respectively. The study from Pakistan by Bari *et al*.(2001) mentioned that nearly 12.5% of cases and 4.5% of controls had history of imprisonment. The statistics of present study make known these figures as 7.6% and 5.3%, respectively.

j) Sharing of Contact Lenses

This factor is first time being introduced as the suspected risk factor of hepatiits C infection in the recednt study. It is presumed that using and sharing contact lenses is generally unacceptable towards the eyes which enabled it to create sever complications like corneal ulcers and lead to infection. Sharing of lenses is common in females compared to males. It is noticed that certain females acquire the lenses from their mates, colleagues, sisters and other relatives on certain occasions to attend wedding ceremony or other celebrations and made the insertion into their eyes. During such behaviour, the infection may be transmitted, if infected leneses are embraced. Most likely the associated risk of sharing contact lenses is considerably least possible, yet that shoule not be ingrored and studied.

The results from present study reflected that out of whole sample, only 28 (2%) patients shared their lenses. Moreover, among the cases and controls the positive history of sharing of lenses is reported as 1.9% and 2.1%, respectively, showing insignificant association with the disease.

k) Minor Surgery by Barber

Kandeel *et al*.,(2012) documented from the Egypt in which roughly 41% of cases acquired cut on the chronic wounds from the barber. In Pakistan also, it is a dilemma that certain barbers perform minor surgical practices like circumcision at their ends. They have absolutely no steralized instruments and undertake cuts through the exact same sharp blade upon several customers. Our data reveal that overall 169 (12.1%) of patients underwent minor surgery or cut from the barber, whilst among the cases and controls this frequncy

allocated as 128(18.3%) and 41 (5.9%), respectively. The counts (percentages) are substantially raised in cases than controls articulating positive association for this factor with hepatitis C infection. Such practices are frequent in rural areas where general Surgeons are miles away and poor people have to rely only on barbers to get the primary relieve. Thus, the present study data showed that percentage of cases having cut from barbers is substanially low as compared to aformentioned Egyptian study.

l) Homelessness and Hostel life

Nyamathi *et al*.(2002) and Sherriff and Mayon-White (2003) mentioned that homelessness is an evident risk factor for hepatitis C infection but an insufficient investigation originated for studying the relationship between living in a hostel and hepatitis C infection. Within hostel environments the risk of HCV infection may increase (Neale and Stevenson, 2012).

This study overall data showed that 268 (19.14%) patients have a history of homelessness and passed hostel life whilst, among the cases and controls this figure distributed as 161 (23%) and 107 (15.29%), respectively. The results intimate about a relationship between hostel life and HCV infection. This factor is also being introduced first time in present study concering Pakistan.

m) Facial barber Shave (Males only)

In Pakistan, barber shave is a very well reported and prominent risk factor among the males. This factor is normally noticed in developing countries like Pakistan, Bangladesh, Egypt, and China. For instance, in Bangladesh, Ashraf *et al*. (2010) found that barber shaving is significant risk factor of HCV with 1.49 folds greater risk. A study from rural Sindh, Pakistan reveals that counts (percentages) of cases and controls for having facial barber shaving are 36.3%, 21.7%, respectively. On comparison, the present statistics indicate that proportions of positive history of this factor are 68.43% and 45.05% among the cases and controls, respectively, which are fairly higher as compared to Bangladesh, Egypt (39%) (Medhat *et al.*, 2002) and the studies from rural Sindh, Pakistan. A cross sectional survey by Janjua and Nizamy (2004) is carried out on the barbers in Rawalpindi/Islamadad cities to know what sort of knowledge and practices they have about the trasnmission of hepatitis C and B virues. The authors observed that majority of the barbers are absolutely ignorant and had habits of reusing of razor for 46% of clients whilst cleaned with antiseptic solution for

only 11.4% (Qureshi *et al.*, 2009). Of the total males only 18% had their facial-shave at home whereas the remaining had by the barbers.

n) Sharing of Razors (Males only)

A study by Mustufa *et al*.(2010) from teachers in Karachi, Pakistan showed that about 13% of teachers shared their razors for shaving purpose. Tanwandee *et al*.(2006) presented a case control study from Thailand and implicted that risk of HCV increases 2.3 folds for the individuals having habit of sharing their used razors. The present study observed about 22.25% of cases had history of sharing of their razors for shaving purpose or removing of un-wanted hairs.

o) Habit of Nails Cutting from Barbershop (Males only)

It is also observed that among the Pakistani males, many possessed the habit of nail cutting from the barber shops, particularly in rural areas. A communal nail clipper is commonly implied at numerous customers that primarily enhances the likelihood of disease. This factor is not really reviewed in previous literature and the researcher felt on personal experience that most of the males frequently practiced nails extraction from the barbers. The current study data indicates that about 269 (36.45%) individuals have a history of nails cutting from barber while 469 (63.55%) have not. Moreover, among the cases and controls 49.35% and 22.38% patients mentioned that they intentionally visited barber's for nails cutting.

p) Removal of Un-Wanted Hair from the "Hammams"(Males only)

In Pakistan, the researcher witnessed many of the people who habitually or intentionally take a bath from the barber shops on payment, locally named as "Hammams". During taking shower these people make use of the razor blade already placed inside the "Hammams" for removing their un-wanted hair. That particular razor blade had been used by the various persons and might transmit the infection into the healthy ones. To the best knowledge of researcher, this factor is also being considered first time as a possible risk factor of hepatitis C among the males of (province) Punjab. The current study statistics showed that among the male patients about 47.76% of male patients took an under shave inside barber's "Hammams". Similary, the counts (percentages) of cases and controls are reported as 245 (62.18%) and 117 (32.14%), respectively indicating an positive association with the disease.

q) Armpit Shave from Barber (Males only)

Bari *et al*.(2001) found that the patients having armpit shave by a barber suggested a positive association with the HCV infection among the adult males of Rawalpindi district, Punjab. They claimed that about 48.2% of cases had positive history of armpit shave from barber's shops. On comparison with recent study, almost similar proportion of cases is identified (40.35%) having habit of armpit shave from the barber. This habit is equally frequent in rural as well as urban population. It is subsequently noticed that barbers are getting awareness through print and electronic media and moving over new blade for each shave but they did not feel comfortable to change a new blade for armpit shave, in particular.

r) Use of Alcohol

The counts (percentages) of cases which disclosed the habit of alcoholic use are reported as 12 (1.7%) while this figure is only 1% in controls. This indicated that use of alcohol has no relationship with the disease in the (province) Punjab, Pakistan. In a case control study, Balasekaran *et al* (1999) reported from Southern United States that frequent use of alcohol has significant association with disease. In China, the positive cases of alcolism is reported as 10.7% (Lin *et al.*, 2003). While Pakistan, being the Islamic Republic of Pakistan, prohibitate the use of alcohol for the Muslims (constitution of Pakistan 1973, article 37-h). Another study by Gheorghe *et al* (2010) is aimed at describing the seroprevalence of HCV infection and its associated risk factors in Romania found insignificant association with the HCV.

s) Local Migration/ Travelling abroad

Viral hepatitis is not really consistently spread within and among the countries, therefore, the individuals who belong to more prevalent regions have greater susceptibility to diease. And the population has not equivalent exposure to HCV. Thus the existing migration and its implications for hepatitis C have been considered. Bollepalli *et al*.(2006) mentioned the significance of travelling inside or outside the country for getting the acute hepatitis C. Another study by Bollepalli *et al*.(2006) reported a risk of infection while travelling through India. Pohjanpelto (1992) reported from Finland that about 8.1% of the patients have lived or travelled in Southern countries, especially in Africa and in the Eastern Mediterranean. The descriptive of present study reveal that about 39.7% of cases has locally migrated from one region to some other region of Pakistan. Similary, 3.7% of cases declared the history of foreign travelling. In Rawalpindi Division, the majority of the patients stated that they had

moved from earth quack regions after 8th October 2005 earth quack in Pakistan. Similary thousands of refugees from Aghanistan are moved to Pakistan after United States-Afghanistan War in 2001. About 1.7 million Afghan refugees migrated to Pakistan and started to live in different provinces of Pakistan. Approximately 25,000 Afghans national are living in a refugee camp between the capital Islamabad and Rawalpindi city. Similarly, in Sargodha Division many cases are identified who are migrated from Karachi and other districts of Pakistan.

t) Family history of Hepatitis

Family history of hepatitis is an other well established risk factor of hepatitis C infection. He *et al*. (2011) documented in a case control study from China about 0.3% of cases and 0.7% of controls, suggesting no significant difference. However, Abbas *et al*. (2008) claimed that family history of hepatitis/liver disease can be an independent variable of HCV infection in patients of rural Sindh, Pakistan. In this study, the counts (percentages) of cases and controls for postive family history of hepatitis are referred to as 16.3% and 14.7%, respectively. When compared, the present study data suggests that these findings are 220 (31.4%) and 106 (15.1%) out of cases and controls. A parallel proportion is observed in control group when compared with aforesaid study by Abbas *et al*. (2008) whereas percentage of positive hitory of hepatitis among the cases is twice compared with Abbas *et al*. (2008). In present study, family history of hepatitis is indirectly inquired from the patients which included information about first order relatives (father, mother, brother, sister and spouses if married) and then analyzed by summing up all infromation in a single variable *i.e* Family history of hepatitis.

u) Family History of Jaundice

A hospital based case control study by Ghaffar *et al* (2009) explored that family history of jaundice is another probable risk factor of HCV among the women of reproductive age, in Quetta, Balochistan (province) of Pakistan. They stated that about 46.8% of the cases and 20.09% of the controls are living or ever lived with jaundice patient in household. In contrast, the present study found a fairly low proportion of cases (5.6%) and controls (6.1%) allied with family history of jaundice.

v) Extra-marital Relation (Male only)

A latest study has shown that marital relation is also the reported risk factor in international literature (Zhao *et al.*, 2013). While in another case control study by Neal *et al*

(1994) this factor found insignificant in UK. Although, this rout of transmission has minor association on hepatitis C infection but still it should be taken into the account for its real impact. Regarding Pakistan it could not be studied due to societal constraints. However, a little effort has been made to inquire the patients about the history of extra-marital marital relation in this study from the male patients only. The descriptive statistics showed that about 47.86% male patients have history of extra-marital marital relation. In univariate analysis this factor is found insignificant and could not be considered in multivariate analysis.

Table 4.1: Univariate Analysis of Socio-Demographic Factors

Variable	Category	Outcome		Total	Chi-Square (p)	Univariate Logistic Regression	p-value
		Controls	Cases			Odds Ratio (95%CI)	
Gender	Male	364(52.0%)	397(57.0%)	761(54.0%)	3.14(0.086)	0.827(0.67-1.02)	0.077
	Female	336(48.0%)	303(43.3%)	639(45.6%)			
Age group	<=34	391(55.9%)	328(46.9%)	719(51.4%)	11.35(0.001)	1.44(1.16-1.77)	
	>34	309(44.1%)	372(53.1%)	681(48.6%)			<0.0001
Marital Status	Never married	197(28.1%)	99(14.1%)	296(21.1%)	41.15(0.000)	2.39(1.82-3.11)	<0.0001
	Ever married	503(71.9%)	601(85.9%)	1104(78.9%)			
Family status	Nuclear	351(50.1%)	315(45.0%)	666(47.6%)	3.71(0.061)	1.23(0.99-1.52)	0.054
	Joint	349(49.9%)	385(55.0%)	734(52.4%)			
Residential locality	Urban	256(36.6%)	296(42.3%)	552(39.4%)	4.79(0.033)	0.787(0.635-0.98)	0.029
	Rural	444(63.4%)	404(57.7%)	848(60.6%)			
House status	Owned	560(80.0%)	541(77.3%)	1101(78.6%)	1.54(0.24)	1.18(0.91-1.52)	0.21
	Rented	140(20.0%)	159(22.7%)	299(21.4%)			
Family size		7.38±3.71	7.53±3.84	7.49±3.76	0.451	1.01(0.98-1.04)	0.451
No. of person sharing the room		3.31±1.65	4.52±3.08	3.91±2.51	0.000	1.26(1.12-1.33)	<0.0001
Patient's Education	No Schooling	216(30.9%)	439(62.7%)	655(46.8%)	190.97(0.000)	Reference	<0.0001
	1-5	90(12.9%)	78(11.1%)	168(12.0%)		10.73(7.25-15.87)	
	6-10	204(29.1%)	147(21.0%)	351(25.1%)		4.57(2.87-7.30)	
	11-16	190(27.1%)	36(5.1%)	226(16.1%)		3.80(2.51-5.76)	
Income group (in Rupees)	<=5000	146(20.9%)	182(26.0%)	328(23.4%)	6.56(0.087)	Reference	0.087
	5000-15000	470(67.1%)	436(62.3%)	906(64.7%)		1.59(0.934-2.71)	
	15000-25000	47(6.7%)	53(7.6%)	100(7.1%)		1.18(0.72-1.92)	
	>25000	37(5.3%)	29(4.1%)	66(4.7%)		1.44(0.77-2.69)	
Father's Education	No	481(68.7%)	568(81.1%)	1049(74.9%)	28.78(0.000)	0.51(0.40-0.64)	<0.0001
	Yes	219(31.3%)	132(18.9%)	351(25.1%)			
Mother's education	No	546(78.0%)	588(84.0%)	1134(81.0%)	8.19(0.005)	0.68(0.52-0.89)	<0.0001
	Yes	154(22.0%)	112(16.0%)	266(19.0%)			

Table 4.2: Univariate Analysis of Risk Factors

Variable	Category	Outcome		Total	Chi-Square(p)	Univariate logistic Regression	P-value
		Controls	Cases			OR (95%CI)	
Family history of Hepatitis	No	594(84.9%)	480(68.6%)	1074(76.7%)	51.96(0.000)	2.57(1.98-3.34)	<0.0001
	Yes	106(15.1%)	220(31.4%)	326(23.3%)			
Patient history of Jaundice	No	599(85.6%)	550(78.6%)	1149(82.1%)	11.66(0.001)	1.61(1.23-2.14)	0.001
	Yes	101(14.4%)	150(21.4%)	251(17.9%)			
Family history of jaundice	No	657(93.9%)	661(94.4%)	1318(94.1%)	0.207(0.733)	0.64(0.57-1.41)	0.649
	Yes	43(6.1%)	39(5.6%)	82(5.9%)			
Family history of liver disease	No	665(95.0%)	585(83.6%)	1250(89.3%)	47.79(0.000)	4.0(2.52-5.54)	<0.0001
	Yes	35(5.0%)	115(16.4%)	150(10.7%)			
History of blood transfusion	No	528(75.4%)	387(55.3%)	915(65.4%)	62.72(0.000)	2.48(1.98-3.12)	<0.0001
	Yes	172(24.6%)	313(44.7%)	485(34.6%)			
History of blood donation	No	579(82.7%)	525(75.0%)	1104(78.9%)	12.49(0.001)	1.60(1.23-2.07)	<0.0001
	Yes	121(17.3%)	175(25.0%)	296(21.1%)			
Dental Surgery	No	531(75.9%)	379(54.1%)	910(65.0%)	72.54(0.000)	2.66(2.12-3.34)	<0.0001
	Yes	169(24.1%)	321(45.9%)	490(35.0%)			
Kidney dialysis	No	700(100.0%)	699(99.9%)	1399(99.9%)	----	----	----
	Yes	0.000	1.000	1.000			
Tattooing	No	642(91.7%)	572(81.7%)	1214(86.7%)	30.38(0.000)	2.48(1.78-3.45)	<0.0001
	Yes	58(8.3%)	128(18.3%)	186(13.3%)			
Body piercing	No	372(53.1%)	373(53.3%)	745(53.2%)	0.003(0.957)	0.99(0.81-1.23)	0.960
	Yes	328(46.9%)	327(46.7%)	655(46.8%)			
Multiple marriages	No	662(94.6%)	651(93.0%)	1313(93.8%)	1.48(0.267)	1.31(0.85-2.03)	0.224
	Yes	38(5.4%)	49(7.0%)	87(6.2%)			
Injected drugs, even once	No	681(97.3%)	650(92.9%)	1331(95.1%)	14.65(0.000)	3.0(1.61-4.73)	<0.0001
	Yes	19(2.7%)	50(7.1%)	69(4.9%)			
History of accidental needle stick	No	507(72.4%)	370(52.9%)	877(62.6%)	57.29(0.000)	2.34(1.88-2.93)	0.000
	Yes	193(27.6%)	330(47.1%)	523(37.4%)			
Sharing of syringes	No	669(95.6%)	661(94.4%)	1330(95.0%)	0.962(0.391)	1.27(0.79-2.07)	0.328
	Yes	31(4.4%)	39(5.6%)	70(5.0%)			
Sharing of tooth brush/Miswak	No	574(82.0%)	516(73.7%)	1090(77.9%)	13.94(0.000)	1.62(1.26-2.10)	<0.0001
	Yes	126(18.0%)	184(26.3%)	310(22.1%)			
Sharing of nail cutter	No	249(35.6%)	144(20.6%)	393(28.1%)	39.00(0.000)	2.13(1.68-2.71)	<0.0001
	Yes	451(64.4%)	556(79.4%)	1007(71.9%)			
Major/Minor Surgery	No	555(79.3%)	420(60.0%)	975(69.6%)	61.58(0.000)	3.0 (2.01-3.24)	<0.0001
	Yes	145(20.7%)	280(40.0%)	425(30.4%)			

History of Angiography/Angioplasty	No	684(97.7%)	683(97.6%)	1367(97.6%)	0.031(0.863)	1.06(0.53-2.12)	0.860
	Yes	16(2.3%)	17(2.4%)	33(2.4%)			
History of endoscope or gastroscope	No	677(96.7%)	616(88.0%)	1293(92.4%)	37.65(0.000)	4.01(2.50-6.45)	<0.0001
	Yes	23(3.3%)	84(12.0%)	107(7.6%)			
Road crash injury	No	576(82.3%)	517(73.9%)	1093(78.1%)	14.52(0.000)	1.64(1.27-2.17)	<0.0001
	Yes	124(17.7%)	183(26.1%)	307(21.9%)			
History of cuts	No	446(63.7%)	304(43.4%)	750(53.6%)	57.91(0.000)	2.29(1.85-2.84)	<0.0001
	Yes	254(36.3%)	396(56.6%)	650(46.4%)			
Hospitalization	No	445(63.6%)	321(45.9%)	766(54.7%)	44.33(0.000)	2.06(1.66-2.55)	<0.0001
	Yes	255(36.4%)	379(54.1%)	634(45.3%)			
History of injections/Intravenous drips	No	280(40.0%)	141(20.1%)	421(30.1%)	65.63(0.000)	2.64(2.08-3.36)	<0.0001
	Yes	420(60.0%)	559(79.9%)	979(69.9%)			
History of branula insertion	No	430(61.4%)	286(40.9%)	716(51.1%)	59.28(0.000)	2.31(1.86-2.86)	<0.0001
	Yes	270(38.6%)	414(59.1%)	684(48.9%)			
History of acupuncture	No	696(99.4%)	698(99.7%)	1394(99.6%)	0.67(0.452)	0.50(0.09-2.73)	0.422
	Yes	4(0.6%)	2(0.3%)	6(0.4%)			
Ever imprisoned	No	663(94.7%)	647(92.4%)	1310(93.6%)	3.04(0.102)	1.47(0.951-2.27)	0.083
	Yes	37(5.3%)	53(7.6%)	90(6.4%)			
Local migration	No	612(87.4%)	422(60.3%)	1034(73.9%)	133.55(0.000)	5(3.50-6.00)	<0.0001
	Yes	88(12.6%)	278(39.7%)	366(26.1%)			
Travelling Abroad	No	682(97.4%)	674(96.3%)	1356(96.9%)	1.50(0.225)	1.46(0.79-2.69)	0.223
	Yes	18(2.6%)	26(3.7%)	44(3.1%)			
Did you use eye connectors, lens etc	No	685(97.9%)	687(98.1%)	1372(98.0%)	0.15(0.849)	0.86(0.41-1.83)	0.703
	Yes	15(2.1%)	13(1.9%)	28(2.0%)			
Minor surgery by a barber	No	659(94.1%)	572(81.7%)	1231(87.9%)	50.94(0.000)	4(2.49-5.20)	<0.0001
	Yes	41(5.9%)	128(18.3%)	169(12.1%)			
Eye surgery	No	677(96.7%)	658(94.0%)	1335(95.4%)	5.82(0.022)	1.88(1.12-3.16)	0.017
	Yes	23(3.3%)	42(6.0%)	65(4.6%)			
Ear wax removal from hospital	No	692(98.9%)	679(97.0%)	1371(97.9%)	5.95(0.023)	2.68(1.18-6.08)	0.019
	Yes	8(1.1%)	21(3.0%)	29(2.1%)			
Homelessness and hostel life	No	593(84.71%)	539(77.0%)	1132(80.86%)	13.46(0.000)	1.67(1.26-2.17)	<0.0001
	Yes	107(15.29%)	161(23.0%)	268(19.14%)			
Alcohol abuse	No				1.33(0.356)	1.73(0.68-4.41)	0.245
	Yes						

4.2 Univariate Analysis

In order to observe the statistical significance of relationship between every single variable and the disease status, t-test, Chi-square and Fisher's Exact tests are applied (where required). A separate univariate logistic regression analysis is performed together with these statistics before moving to multivariate anslysis. It is also observed that the Chi-square test and univariate logistic regression provide identical results. However, Chi-square test is theoretically more of descripitve test whilst regression analysis enables us to estimate a predictive capacity of the factor. Furthermore, a univariate logistic regression makes it possible to evaluate crude odds ratios with their 95% CIs and describe the nature of relationship between the independent and the outcome variables that could not be assessed by Chi-square test solely. The Chi-square test simply explains the association between exposure and the outcome nevertheless signifies the sign of relationship. All *p*-values are 2-tailed and the p-value<0.05 is taken as significant. And the varibles having p-value<0.20 are selected for final multivariate anlaysis. The details of Univariate analysis are explicated in **Table 4.1** and **Table 4.2**.

From **Table 4.1,** among the socio-demographic factors, analysis showed that patients age, marital status, residential area, No.of persons sharing the room, patients's education, father and mother's education are the significant factors at 5% level of significane. While other variables such as gender, family status, house status, family size, household income are insignificant (*p*-value>0.05). The crude odds ratio of each respective variable is also given and an odds ratio greater than 1 indicates a positive association of that particular factor with the disease and vice versa. Among the list, patient, father and mother's eduction are strongly but negatively associated with the disease indicating that these are protective factors against the disase. The household income is categorized into four groups which is found to be insigificant at p-value<0.05. But it is still considered in multivariate analysis as its p-value is less than 0.20. however, it is noticed that majority of the patients belong to poor families. In contrast, the factors showing positive association with the disease are the actual risk factors.

A case control study by Bari *et al*.(2001) from Rawalpindi-Islamabd among the adult males found that age>35 years and marital status are the significant factors in their univarite logistic regression analysis. The crude odds ratio (95% CI) of age group and marital status of present study [1.44(1.16-1.77),2.39(1.82-3.11)] are slightly lower as compared to Bari *et al*.(2001) study *i.e* [4.2 (1.5±11.8), 5.5(2.4±12.7)], respectivley. However, the figures relecting in present sample are more reliable because of large sample size and inclusion of

female patients as well. Our findings are supported by Akhtar *et al*.(2004) describing identical crude odds ratios (95% CI) for the age and marital status. Another study by Abbas *et al*. (2008) also suggested that age>34 years and gender are the potential factors in a rural Sindh, Pakistan by the univariate logistic analysis. While in present study, gender is refelcted as insignificant factor. Moreover, Abbas *et al*. (2008) reported that "No.of sharing the room" is insignificant in their sample but present study reveals this factor as a significant factor in univariate anlaysis.

Another importatn factor i.e Patients' education is also negatively associated with the disease which indicats that with the rise in education level of the patient, the risk of disease reduces from 10.73 to 3.80. Zaller *et al* (2004) reported contradictory finding from Georgia, USA that patient's education is an insignificant factor but still they considered it as an important factor. However, Ghaffar *et al*.(2009) established a significant relationship between the education and disease status which support to this study finding also. Importantly, the present study ensured the importance of parents' education that is unfortunately missing in literature and could not be studied in the similar context.

From **Table 4.2**, separate analysis of medical, behavioral and family history related risk factors are also examined in the univariate analysis. The results suggested that family history of hepatitis C, patient history of jaundice, family history of liver disease, blood transfusion, blood donation, dental surgery, accidental needle stick, major/minor surgery, endoscopy, history of cuts, hospitalization, history of injections, history of branula insertion, acupuncture, eye surgery are the significant risk factors of HCV under Univariate analysis. In addition, some more factors from the behavioral characteristics and personal choices of patients have positive association with the hepatitis C infection i.e tattooing, body piercing, multiple marriages, Injected drug use, sharing of toothbrushes/miswak, nail cutter sharing, road crash injury, imprisonment, local migration, minor surgery by the barber, ear wax removal, homelessness and hostel life.

In present study, family history of hepatitis C is a significant risk factor of HCV with a crude odds ratio (OR=3.0) which explains that the risk of disease increases 3 times in patients with its positive history than others. A study by Abbas *et al*.(2008) from rural Sindh, Pakistan and another parallel study from Korean population suggested that family history of hepatitis is insignificant under the univariate logistic regression. However, another identical study showed a positive association this particular risk factor. A matching crude odds ratio is identified by this study (OR=6.41).

Further , Bari *et al*. (2001) performed a case control study on adult males in Rawalpindi-Islamabad population and univariate analysis suggested that the corresponding odds ratios of patient history of jaundice, blood transfusion, dental surgery, major/minor surgery**,** hospitalization and tattooing are 3.2, 3.6, 1.7, 3.4, 4.1, 0.9, respectively. Whilst the present study demonstrates the comparable odds ratios for the same risk factors as 1.6, 2.5, 2.7, 3.0, 2.1 and 2.5, respectively. Other significant risk factors such as history of injections (OR=1.77, 95% CI: 0.69–4.59), blood donation (OR= 3, 95% CI: 0.99–9.09) are reported by Liu *et al*. (2009). Tattooing and ear piercing are insignificant in rural Sindh, population (Abbas *et al.,* 2008).

It is really worth important to describe that every variable possessing p-value<0.05 in univariate analysis does not necessarily mean that they are ultimate risk factors of the study. For this purpose, a multivariate anlysis ought to be performed well before finalizing the results. On the other hand, it may happen that a variable being an insignificant varible in the univariate analysis would develop significant recognition in the final multivariate analysis. Therefore, Hosmer and Lemeshow (2000) and Bari *et al*.,(2001) stated that every variable having p-value<0.20 in univariate analysis should preferably include in multivariate analysis in order to avoid an ultimate loss of some important variables. Hence, the detailed discussion on significant risk factors of hepatitis C would be discussed after carrying out multivariate analysis.

4.3 Multivariate Logistic Regression Analysis

Initially, total 46 factors are considered for model building on overall data set from the possible socio-demographic, clinical risk factors, Family history and behavioral characteristics. Later on, multivariate LR model is performed with 33 predictors which are selected in univariate analysis. As it has been mentioned earlier that the primary objective of univariate analysis is to scrutinize the potential variables for their consideration into final multivariate anlysis. All variables having p-value less than 0.20 are extracted in the univariate analysis and considered for final multiple logistic regression analysis.

It might not be good enough to develop a regression model without adopting its proper procedures and methods. This looks quite unjustifiable to assume that a modeling of LR model is merely simple and straightforward. Actually the researchers often have to put an efficient time, energy and efforts to achieve a parsimonious model using proper statistical procedure. In simple words, the model building could hardly be at once and necessitate

reiterate the model by inclusion or exclusion of certain variables. Model performance should be assessed on every trial. Moreover, the checking of multicollinearity and outliers' detection is also ascertained in this analysis. Following the methodology/procedure given in Chapter No.3, different LR models are run on overall data, males/females data considering urban/rural settings of the patients and northern/southern areas of province Punjab to identify the pertinent risk factors of HCV at their exact place.

4.3.1 Measuring of Multicollinearity

The existence of multicollinearity brings about large standard errors of resulting regression coefficients and makes the findings questionable. Hence, to circumvent the problem and obtain reliable results, checking of multicollinearity was made after performing multivariate analysis per prescribed recommended criterion of Allison (1999) and Menard (2002). From Table 4.4, the results showed that all variables have Variance Inflation Factor (VIF<2.0) and Tolerance limit greater than 0.5 indicating as such no issue of multicollinearity and all covariates are independent.

4.3.2 Identification of Outliers

Before finalizing the suitable model, diagnostics of outlier values can be quite a fruitful assessment for the parsimonious model. Outliers are essentially those observations that are surprisingly far away in the outstanding data observations. These values could be the cause of a measurement error or perhaps an extreme manifestation of natural variability. Such values desired a careful attention of the researcher. Parallel to multiple linear regression modeling, outlier identification should also be the necessary part of the LR modeling but unfortunately its application in epidemiological studies is lacking in literature.

This can be done with the help of graphical representation of residuals as well as predicted probabilities of the fitted model. These graphs are generally known as indexed plots. For instance, on recent study data, certain plots are presented below to uncover outlying observation in the data, (i) predicted probabilities of the fitted model are plotted against each data observation, (ii) different types of residuals are also plotted against the sampled observations, (iii) Cook's influence statistics are plotted against the sampled observations to identify the influential observation (if any). Then finally, residuals are plotted against the predicted probabilities. These graphs are presented in **Figure 4.1-Figure 4.6: Graph of Leverage values against Case ID**

These graphs suggest that case number 499, 637 and 304 would be the points which lead to abnormally large residuals. These are the suspected observations which required particular attention of the researcher. Finally, these are discarded and revised model obtained. On eliminating, a certain decrease in deviance is observed and performance of the model is improved somehow. In our previous study (Ghias and Pervaiz, 2009b) it had been declared how the presence of outliers in the logistic regression model can establish uncertainly large standard errors and incorrect sign of estimated coefficients. Further, inclusion of some unimportant risk factors could also be the part of the model, if outliers not removed in the data. The use of influence diagnostics is also discussed in by Cox & Snell (1989) and Hosmer and Lemeshow (1989). Collett (2002) also explained that an influential observation could be a value that substantially change to model summary and parameters both. He further added that "an outlier may also be influential observation but an influential observation need not necessarily be an outlier". To investigate the influential observations, the Cook's influence statistics and leverage values are presented along with **Figure 4.4** and **Figure 4.5**, respectively and no influential observation is detected because all values of the Cook's statistics are less than 1.

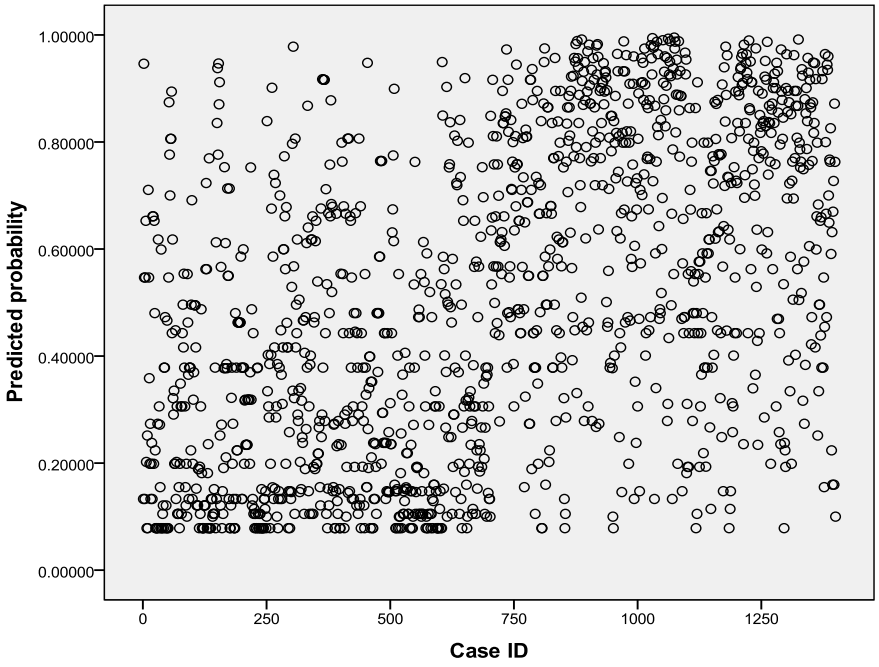


Figure 4.1: Graph of Predicted Probabilities against Case ID

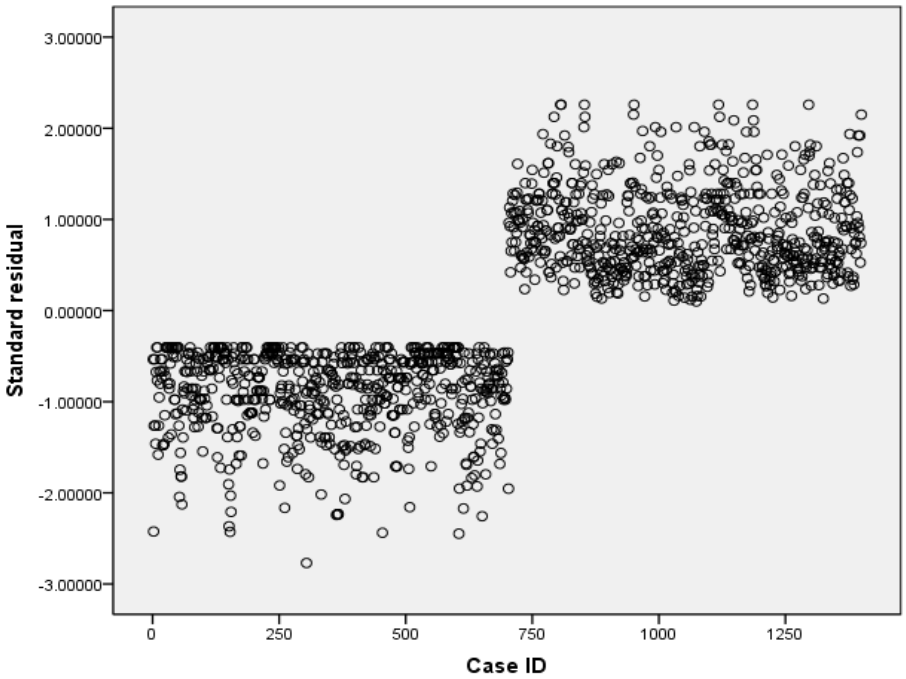


Figure 4.2: Graph of Standardized Residuals against Case ID

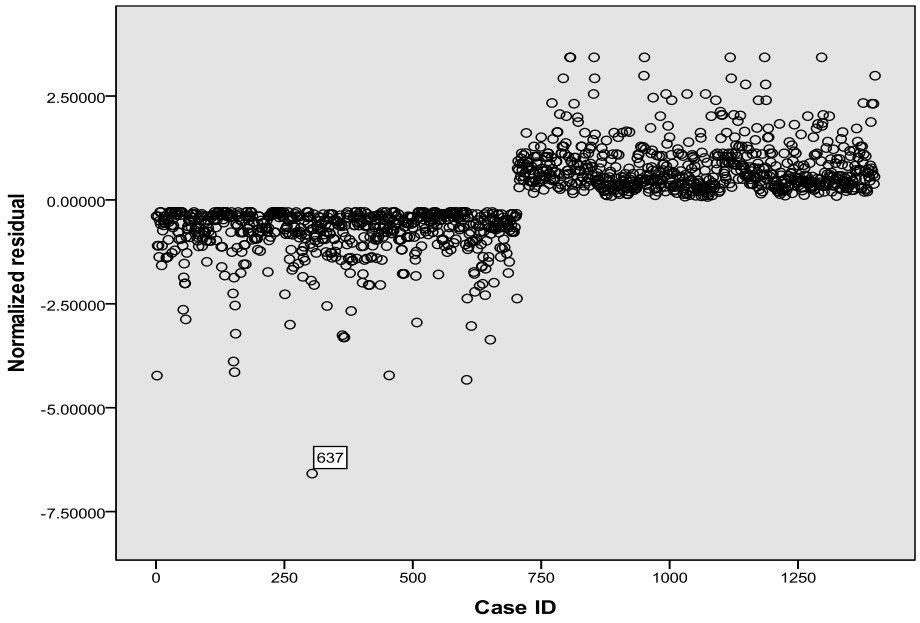


Figure 4.3: Graph of Normalized Residuals against Case ID

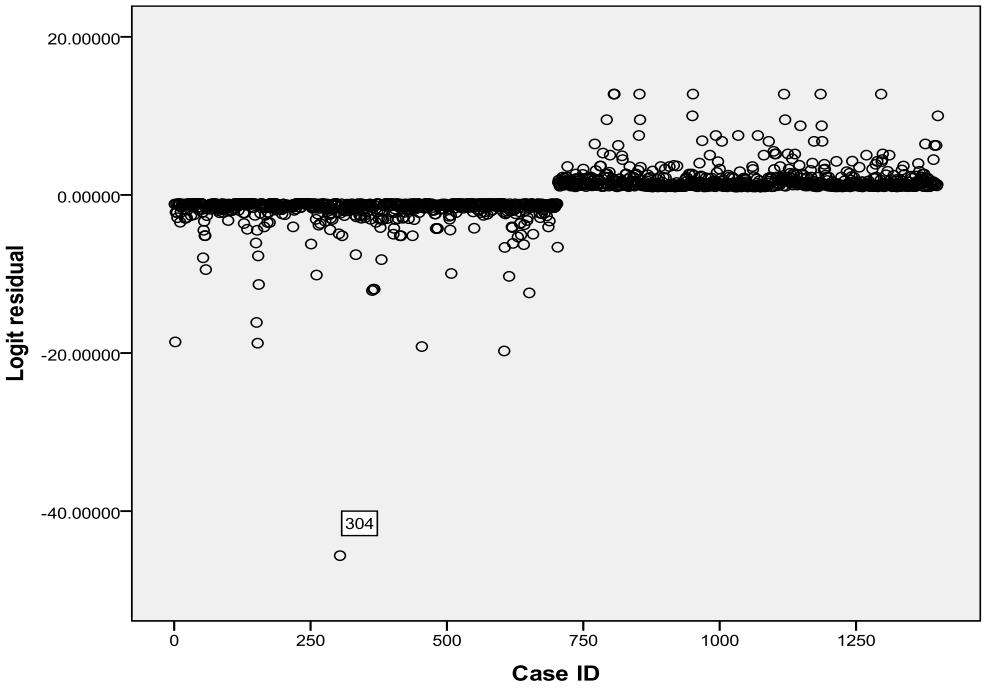


Figure 4.4: Graph of Logit Residuals against Case ID

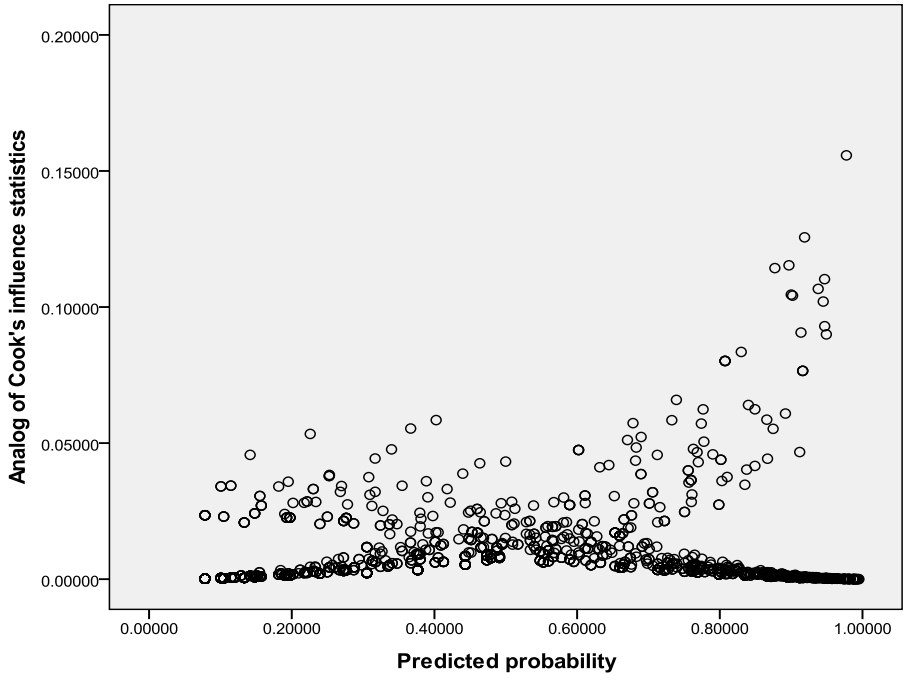


Figure 4.5: Graph of Cook's Influential Statistics against Predicted Probabilities

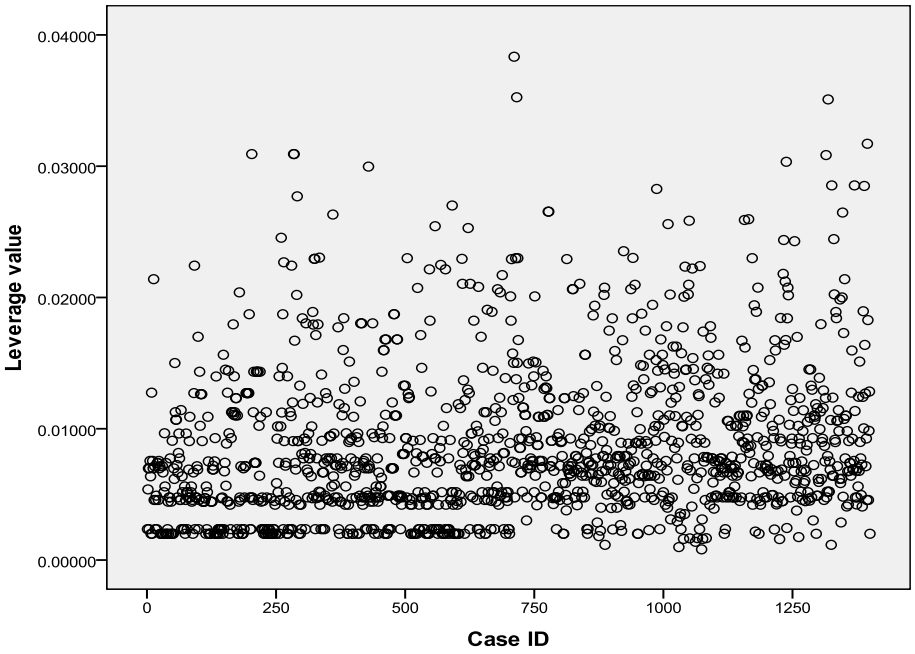


Figure 4.6: Graph of Leverage values against Case ID

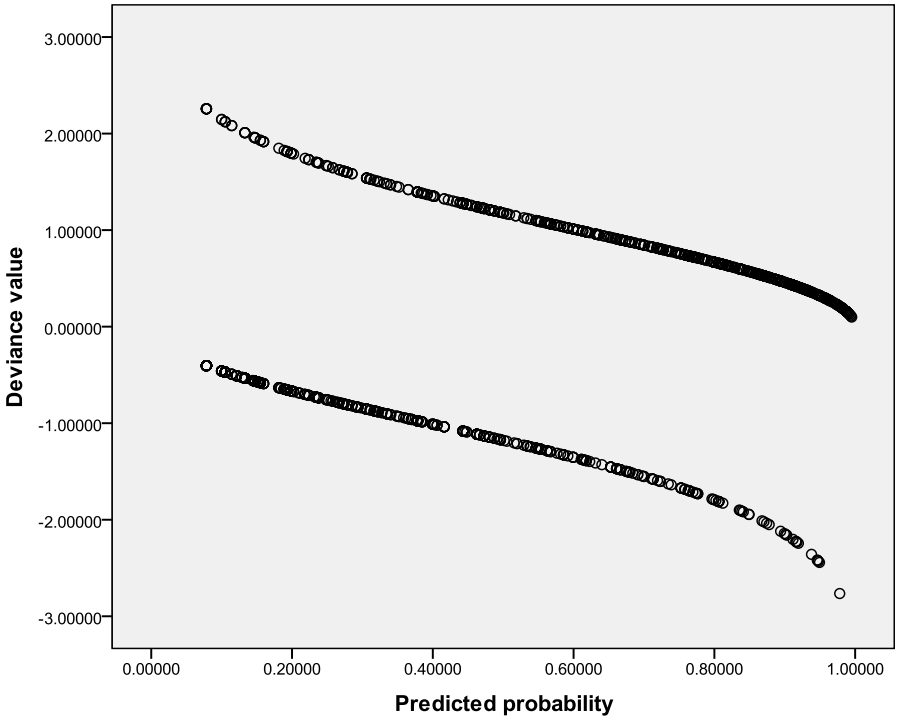


Figure 4.7: Graph of Deviance Values against Predicted Probabilities

4.3.3 Fitting of Multiple Logistic Regression

In most of the epidemiological studies main objective to build a regression model is to establish an exposure–disease relationship. Therefore, researchers are usually more interested to reach on a best fitted model to the data. In this study, to attain the best possible model and also selection of the potential risk factors of HCV infection, several multiple logistic regression models are run on overall data. But, reported only that model which possessed good model performance. Each time, the model performance is examined through different statistical parameters that include -2LL and ROC curve. In the final multiple logistic regression model, the adequacy of the overall fitted model is assessed by the Omnibus and Hosmer and Lemeshow (HL) tests followed by the ROC curve analysis.

The Omnibus test with $\chi^2 = 543.269$ and p-value<0.0001 is found significant which suggests that the model is adequately fitting to the observed data or at-least one of the factors is affecting the outcome variable significantly. Another important alternative is the HL test which can also be implied to examine goodness of the fit of the model. In present data, HL test found to be insignificant with $\chi^2 = 11.729$ at p-value=0.164. While an insignificant HL test depicts good fitted model as the hypotheses to be test are ($H_0 = Model\ is\ best\ fit$; $H_1 = Model\ is\ not\ fit$). Hence, the p-value>0.05 indicates that H_0 is not rejected which means that our fitted model is a good fit (Hosmer and Lemeshow, 2000). Omnibus and HL tests explain about overall significance of the fitted model nevertheless individual explanatory variable. Wald test is the suitable statistics for examining the significance of each individual explanatory variable **Table 4.5**. A step-wise regression is run using Forward likelihood ratio criterion for the final selection of variables in the model. Estimation of logistic regression parameters is based on maximum likelihood method which is obtained through maximizing the joint probability.

Other measures of goodness of the fit include the Cox and Snell R^2 and Nagelkerke's R^2. These measures have revealed their corresponding values as 32.2% and 43.0%, respectively. Importantly, these are analogous to $\boldsymbol{R^2}$ which is broadly used in linear regression to explain how much percentage variation of the model is explained by the explanatory variables. In logistic regression, these are known as Pseudo R^2. It is also pointed out that Nagelkerke's R^2 always reflects higher value as compared to Cox and Snell R^2 (Peat and Barton, 2008).

Rather than using goodness of fit statistics, a presentation of classification table wherein percentage correct classifications of cases and controls are examined. This table provides a useful information regarding performance of the fitted model. Table 4.3 explains that the overall percentage of correct classification is 75.5% whereas, for controls and cases the figures are noted as 77% and 74%, respectively. Among the controls, 538 patients are correctly classified as controls nevertheless 162 patients misclassified as cases. Similarly among the cases, 516 patients are correctly classified as cases and 181 patients are misclassified as controls. Thus, the overall percentage correct prediction of the model is very good.

Table 4.3: Classification Table of Observed and Predicted Outcome

Observed		Predicted		
		Outcome		Percentage Correct
		Controls	Cases	
Outcome	Controls	538	162	76.9
	Cases	181	516	74.0
Overall Percentage				75.4

4.3.4 Receiver's Operating Characteristic (ROC) Curve

From **Figure 4.8** the overall discriminatory power of the logistic regression model to differentiate between cases and controls is also assessed by the use of area under the receiver operating characteristic (AUROC) curve. It is a useful way to graph between (1-specificity) and sensitivity and gives the predictive performance of the fitted model. A curve closer to the upper left diagonal gives better performance. The present analysis showed that the total area under the receiver operating curve is 83.8%, indicating that the model would have a high chance of scoring a case higher risk for disease than control if they are randomly selected from the study sample. Finally, it may be affirmed that the LR model thus developed is adequately fitted to the data.

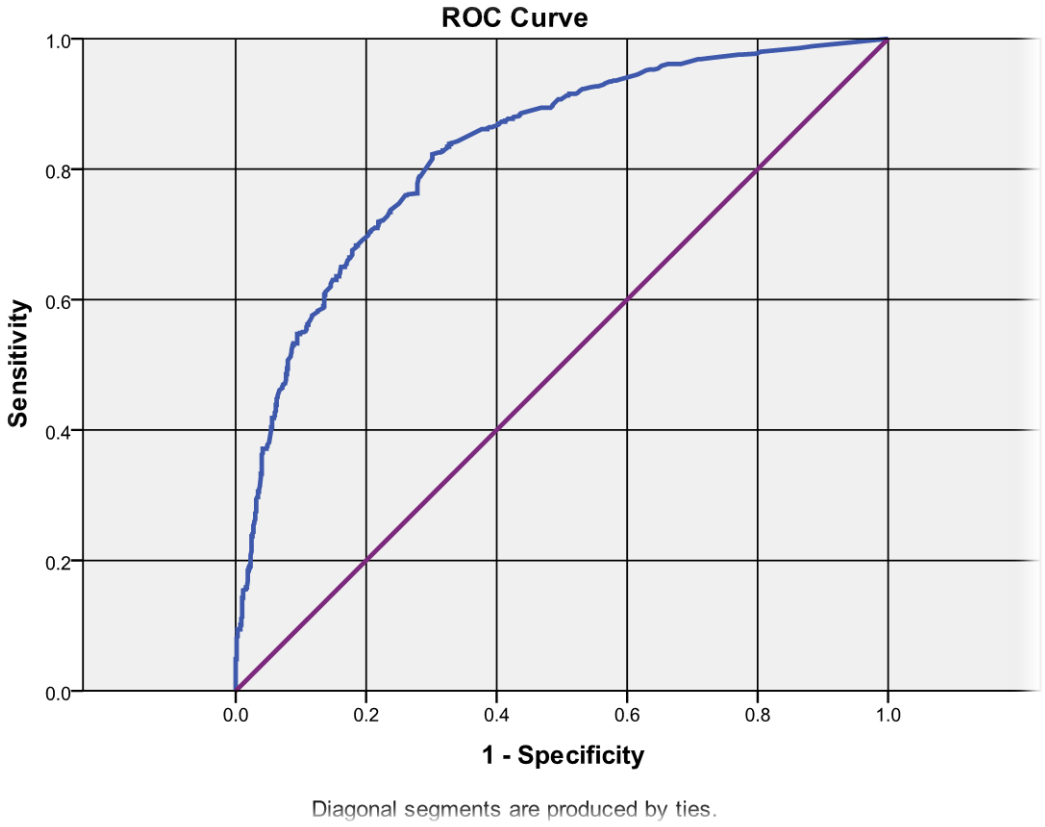


Figure 4.8: Receiver Operating Curve (ROC)

Table 4.5 gives the final output from multivariate logistic regression wherein information of potential risk factors, regression coefficients, Odds ratios along with their corresponding 95% CI's are presented. This table demonstrates that out of all possible socio-demographic, medical history, personal living & behavior and family history related variables only 11 are finally selected in the multivariate analysis. These potential risk factors are summarized as patients' low education, local migration, family history of liver disease, Endoscopy, Family history of hepatitis C, Tattooing, Blood transfusion, Minor surgery by barber, Dental Surgery, General Surgery and History of injections/intravenous drips. It is evident from the regression coefficients sign that all variables have positive association with the HCV infection except one variable that is "patients' low education" which is negatively associated indeed.

Table 4.4: Checking of Multicollinearity after the Multivariate Analysis

Variable	Tolerance	VIF
Patient Education	0.941	1.063
Foreign migration	0.984	1.016
Family history of liver disease	0.977	1.024
Family history of Hepatitis	0.938	1.066
Tattooing	0.982	1.018
History of endoscope	0.974	1.027
History of blood transfusion	0.971	1.03
Dental Surgery	0.935	1.07
Major/Minor Surgery	0.924	1.082

4.3.5 Interpretation of the Model

This section is focused on proper interpretation of each significant risk factor in the fitted model logistic regression model **Table 4.5**. This includes interpretations of regression coefficients, odds ratios and their associated 95% CI's. A systematic discussion of each significant risk factor is also entailed making comparison with other similar studies from national or international literature.

a) Patient's Education

In this study, it is previously described that around 62.7% of cases are absolutely have no formal school education, in the province, Punjab. This factor is also significant in Univariate logistic regression analysis and now appeared in multivariate analysis depicting its real importance in controlling and promoting HCV infection among the general masses. From **Table 4.5** it is evident that patient's education has negative association with the disease indicating that as well as patients education improves the risk of hepatitis C infection decreases. The corresponding estimated odds ratio and 95% CI interval are found as (Adj OR= 0.196; 95% CI: 0.150-0.254) means that about 80.9% (1-0.196=0.804) risk of disease reduces in patients who had well education compared to others. On the other hand, the subjects who have undergone formal education are less likely to be HCV positive. Akhtar *et al*.(2004) presented a case control study from male volunteer blood donors in Karachi and found that education is the significant risk factor of HCV in multivariate analysis. Another study from low-risk country (Finland) reported that patients who have poor education are at

higher risk (Adj. OR= 0.3, $p < 0.001$). In contrast, a study by Zaller *et al* (2004) is presented among blood donors in Georgia, USA to evaluate risk factors of hepatitis C infection and suggested that patient's education has no significant association with the disease. The some other studies by Shazi and Abbas (2006), He *et al*. (2011) and Wolff *et al* (2008) showed the consistent findings with the present study and ascertained that low level of education is an established risk factor of HCV in Pakistan, China and Brazil, respectively.

b) Local Migration/Travelling

It is well documented that prevalence of HCV differs within and between countries substantially and also the people who have lived with more prevalent regions of HCV and moved to different regions are expected to spread the infection. In this study history of local migration is also found to be a significant risk factor in multivariate logistic regression. This factor is also appeared to be significant in univariate analysis revealing the highest odds ratio compared to other stated risk factors. Bollepalli *et al*.(2006) mentioned that travelling of individuals inside or outside the country could be a suspected risk of the acute hepatitis C. In this study, the odds ratio of local migration along with 95% CI is reported as (Adj OR=2.777; 95% CI: 2.01-3.85) suggesting that risk of disease is 3 times higher in patients who had history of local migration/travelling controlling for other significant covariates.

During data collection certain reasons of migration are also inspected, firstly, in Pakistan contemporary urbanization is rapidly growing and families are generally migrating away from rural areas to urban dwellings for better livelihood and their children's education. The researcher observed that in Rawalpindi division, many of the patients expressed that they have migrated out of earth quack regions immediately after 8th October 2005 earth quack in Pakistan. One more rationale of migration is viewed that thousands of afghan refugees are found to be migrated Pakistan soon after Afghan war in 2001and also dispersed around many regions of Pakistan. Unfortunately, Pakistan is also confronting with big challenge of terrorism and Pakistan Army is combating against terrorism in many remote areas of Pakistan. Therefore, lot of people have been migrated even so from the other provinces to Punjab which is regarded as fairly better secure province in respect of terrorism. It is noticed that there is no proper blood screening arrangements has been made for the migrants by the government to diagnose HCV in migrants. To the best comprehension of the researcher, this aspect is first time has been highlighted in this study with regards to hepatitis C and need to cope with the circumstances seriously.

c) Family History of Liver Disease

Another potential risk factor of HCV identified in multivariate logistic regression is "family history of liver disease". The adjusted odds ratio described that the risk of having HCV is increased 3 times in patients who has positive family history of liver disease compared to others, controlling for other significant covariates. Abbas *et al*.(2008) supported this study and also suggested that family history of liver disease is an independent risk factor of HCV infection in rural Sindh, Pakistan. Another study by (Qureshi *et al.*, 2009) from Pakistan reveals that 4.3% of cases reported any one family member died with liver disease ; OR=1.4; 95% CI (0.7-2.5). In contrast, a similar study by Yabuuchi *et al*. (1993) from Japan reported that factor under discussion is insignificant.

d) History of Endoscopy/Gastro scope

Another risk factor in the present study is a history of undergoing endoscopy. The strongest association is observed in those patients who reported a history of undergoing either an endoscopy or gastro scope. The corresponding adjusted odds ratio and the 95% CI are observed as 3.36 (1.34-4.15). Whilst in our recent paper (Ghias et al., 2012a), the associated risk of undergoing endoscopy is reported as (OR=4.85, 95% CI:1.03-22.81) in patients belonging to Gujranwala district, Pakistan. This association concurs with another study, also from Pakistan, which found that endoscopy is a significant risk factor among women (Hashmi *et al.*, 2010). Researchers in France have reported that nosocomial HCV infection may occur from exposure to contaminated equipment (González-Candelas *et al.*, 2010). A prospective cohort study from Egypt, the highest HCV prevalent country in the world, in which 859 participants are recruited, 149 of whom underwent an endoscopic procedure during follow up, reported that 2 patients seroconvert to HCV positive within three months (Mikhail *et al.*, 2007). Transmission of HBV infection by endoscopy has been reported, but it occurs infrequently (Birnie *et al.*, 1983). However, there are no reports, even in Pakistan, to our knowledge, of other blood borne infections, such as HIV, being transmitted by endoscopy. Endoscopy as a risk factor for HCV is not confined to developing countries. A case-control study, carried out in France (n=1207 including 450 cases) reported that a history of undergoing endoscopy has a significant association with HCV infection, with an estimated odds ratio of 1.9 (Karmochkine et al., 2006b). Another French case-control study reported an odds ratio of 8.0 (95% CI= 2.3–27.2) (Delarocque-Astagneau *et al.*, 2007). A conflicting result is reported by Kim *et al*.(1996) in Korean population that endoscopy is insignificant. In present study, endoscopy might have been a part of the diagnostic or prognostic clinical care

of HCV infection, so that reverse causation may be an alternative explanation for this finding. But research from longitudinal studies in other developing countries, supports the assertion that this procedure has its direct association with infection.

e) Family History of Hepatitis C

Family history of hepatitis C is also investigated as another significant risk factor of HCV showing adjusted odds ratio 2.18 and corresponding 95% CI: 1.58-2.99. This study reveals that the patients having hepatitis C in their family members (first degree) have 2.18 times greater risk of HCV as compared to those who have not in their family history. The results are supported by He *et al* (2011) and suggested that family history of hepatitis C is also an evident factor of HCV in our neighboring country, China. However, two parallel studies, one from Pakistan (OR=1.6, 95% CI: 0.9- 2.8) and other representing Korean population stated dissimilar finding (Kim *et al.*, 1996, Abbas *et al.*, 2008).

f) Tattooing

The habit of tattooing is more frequent in western countries, however, some studies inside or other neighboring countries of Pakistan had also established its association with the HCV. In present analysis, tattooing is an evident risk factor of HCV with odds ratio 2.18 (95% CI:1.44-3.29) indicating that the likelihood of disease enhances 2 folds if perhaps people have habit of tattooing. In Pakistan, tattooing is solely in adult men even whilst western countries are possessing in both genders. Despite the fact, really distinctive feel in European countries is noticed because tattooing is done by professionals only but this is simply not the truth with Pakistan. Neal *et al* (1994) presented a case control study of blood donors in the Trent Region (UK) and also identified that tattooing is a significant risk factor of HCV in multivariate logistic regression analysis. Balasekaran *et al* (1999) performed an age-gender and ethnicity matched case control study to examine some new risk factors transmission for sporadic HCV infection in Southwestern United State and suggested "Tattooing" is significant risk factor (OR=5.9, 95% CI: 1.8–320.7). Other similar studies from Canada, United States and Texas–Mexico had proven the association of tattooing with HCV infection with odds ratio 5.7 and 2.93, respectively (Delage *et al.*, 1999, Briggs *et al.*, 2001, Hand and Vasquez, 2005).

Similarly, in a cross sectional survey of our neighboring countries China and Iran "Tattooing" is also a significant risk factor of HCV in multiple logistic regression model (Zakizad *et al.*, 2009, He *et al.*, 2011). He *et al*.(2011) further added that tattooing is

considerably more frequent (30.5%) in cases than controls (5.6%). In comparison, a study from Islamabad, Pakistan by Bari *et al*.(2001) exposed that roughly 8.9% of cases practiced tattooing whereas another identical study by Akhtar *et al*.(2004) from Karachi reflects this figure as 7.4%. However, Ahmed *et al*.(2012) reported in their recent community based study from province (Balochistan) , Pakistan that "Tattooing" is insignificant. Some other studies carried out from, Korea, Egypt, and North Florida and South Georgia have not found its positive association with HCV infection (Kim *et al.*, 1996, Medhat *et al.*, 2002, Mishra *et al.*, 2003).

g) Blood Transfusion

This study found that receipt of a blood transfusion is most strongly associated with HCV infection introducing odds ratio and 95% CI as (Adj. OR= 2.04, 95% CI: 1.44-3.29). Having a blood transfusion is other potential risk factor for infection, and this finding is consistent with other studies carried out in Pakistan (Ghias and Pervaiz, 2009c, Idrees et al., 2008). Other developed countries have markedly reduced the risk of blood transfusion (Dwyre *et al.*, 2011), but the present data indicated this procedure as an ongoing continuing source of infection due to poor blood screening system in Province (Punjab) and even in overall Pakistan . In a systematic review of hepatitis C virus epidemiology, Sievert *et al*. (2011) found that this risk factor has also been identified as a major risk factor for hepatitis C infection in China, India, Egypt, Japan, and Korea suggesting for poor screening system in these countries.

h) Minor Surgery by Barber

In this study, minor surgery by a barber, for example, circumcision, ingrown toenail surgery or drainage of abscesses is another potential risk factor of hepatitis C in province (Punjab); where about 12.1% patients underwent minor surgery by the barber. Another study by Wazir *et al*. (2008) done in rural areas of Punjab found that about 58% of barbers are involved in such malicious practices. A study from Egypt is also showed its significance with the disease indicating roughly 41% of cases received cut on the chronic wounds from the barber (Kandeel *et al.*, 2012). In our recent study it was reported that about 13.4% cases intended to get minor surgeries from the barbers in Gujranwala District of Punjab (Ghias and Pervaiz, 2009a). It looks more disgusting when the barber performs minor surgeries at their own level. However, this trend needs to be stopped by spreading proper awareness among the general masses.

i) Dental Surgery

In this study, another well established risk factor of HCV is the dental surgery and has positive association with HCV infection with odds ratio 1.76 (95% CI: 1.25-2.21). This suggests that risk of disease increases about 1.76 folds with the history of dental extraction/surgery, controlling for other significant covariates. A hospital based case control study is carried out by Mele *et al* (1994) from Italy and suggested that dental surgery is the significant risk factor in their multivariate analysis at 5% level of significance. A similar study from Hawaii is also claimed with the supportive conclusion (Lasher *et al.*, 2005). From Pakistan also a hospital based case control study is carried out by Qureshi *et al* (2009) from Karachi, Sindh province of Pakistan and imply with same concordant finding. Moreover, Abbas *et al*.(2008) and Qureshi *et al*.(2009) identified support our finding that dental surgery is the major exposure of HCV infection in Pakistan with odds ratio 2.1 (95%CI:1.1-4.1, p=0.027) and 2.3 (95%CI:1.8-3.0, p-0.001) respectively. Qureshi *et al*.(2009) further added that history of dental filling, extraction and scaling is significantly higher (38%) in male cases than controls (20.9%) in Karachi, Pakistan.

j) General Surgery

Another prominent risk factor of hepatitis C in Pakistan is the "general surgery" and also found significant in present study. The finding suggested that risk of disease increases by 2 folds in patients who have history of general surgery or surgical operations in past as compared to others who have not any history of general surgery. A very comprihensive community based study is carried out by Habib *et al* (2001) and suggested that this is the potential risk factor in the Nile Delta region, Egypt, the most prevalent region for HCV infection. Another study from Iran, out of the common and known risk factors of HCV infection among addicted prisoners, invasive surgery is also pertinent risk factor of HCV (Zakizad *et al.*, 2009). A different case control study from Poland also concurs and reveal analogous finding by identifying minor surgery (OR = 3.2, 95%CI = 1.5-6.7) as the potential risk factor of HCV.

From Pakistan, Idrees and Riazuddin (2008) suggested that about 70% of the cases are obtained through reuse of syringes and general surgery. This situation is quite alarming and demand for effective sterlization and infection control system in hosptials. Bari *et al*.(2001) from male adults in Rawalpindi/Islamabad identified that roughly 42.1% of cases have positive history of major/minor surgery in past. They also explained that this factor is

significant in Univariate logistic regression (OR=3.4, 95% CI:1.8-6.4) but insignificant in multivariate analysis. Qureshi *et al* (2009) from Karachi also showed concordant findings with present study anlysis.

k) History of Injections/Intravenous Drips

The commonest parenteral risk factors are generally the injection therapy and blood transfusion (Jafri and Subhan, 2010) predominantly in developing countries. Kermode (2004) estimated that at least 12 billion syringes are sold every year and hundreds of these are given by un-sterilized syringes. Pakistan is the leading country in which superfluous therapeutic injections are on top and an average of 13.6 injections are received every year per person (Janjua et al., 2006b). Darwish *et al* (1993) and Mohamed *et al* (1996) investigated that the "therapeutic injection" is the common risk factor in Egyptian blood donors' population. A case control study by Comandini *et al* (1998) is completed for identification of sporadic risk factors of HCV infection in a selected sample of hospital in general population in Rome, Italy and history of injection is appeared as an independent risk factor. In addition, several other similar studies have established the positive association of this risk factor with HCV in different regions of the world, for example, Merle *et al* (1999) in general population; Liu *et al* (2009) in a Henan province; Gheorghe *et al* (2010) in Romania are identified the comparable findings in France, China and Romania, respectively.

Janjua *et al* (2005) in a population based cross sectional survey to estimate annual number of injections per person and their associated cost in province Sindh of Pakistan. It is noticed that about 76% patients have the history of injections for curative purpose received from the dispensors and 67% from the qualified general practitioner. In the same way, Bari *et al* (2001) and Shazi and Abbas (2006) supported this study from Pakistan and introduced odds ratio comparable to this study (OR=2.8, 95% CI: 1.1-7.1), respectively.

Table 4.5: Output from Multiple Logistic Regression Model for Overall Data

(Regression Coefficients, Odds Ratio and Their 95% CI's)

Variables	B	S.E.	Wald	P-Value	Adjusted Odds Ratio	95% C.I.for Odds Ratio	
						Lower	Upper
Patient Education	-1.632	.134	148.709	.000	.196	.150	.254
Local migration/Travelling	1.021	.166	37.756	.000	2.777	2.005	3.846
Family history of liver disease	.973	.237	16.843	.000	2.646	1.663	4.212
Endoscopy	.857	.288	8.849	.003	2.357	1.340	4.147
Family history of hepatitis C	.777	.162	23.016	.000	2.176	1.584	2.989
Tattooing	.777	.211	13.611	.000	2.175	1.439	3.286
Blood transfusion	.715	.139	26.302	.000	2.043	1.555	2.685
Minor surgery by barber	.553	.236	5.485	.019	1.739	1.094	2.762
Dental Surgery	.508	.144	12.368	.000	1.661	1.252	2.205
General Surgery	.456	.147	9.564	.002	1.577	1.182	2.105
History of injections/intravenous drips	.418	.150	7.821	.005	1.519	1.133	2.037
Constant	-.817	.156	27.427	.000	.442		

4.4 Gender and Area-specific Logistic Regression Models

Once the analysis of risk factors of HCV is completed in overall data, a separate analysis has now been brought in for men and women separately, keeping in view their differences in biological, lifestyle/behavioral and medical history related factors. Although a number of studies had been under taken for analyze the risk factors of HCV in both males (Akhtar et al., 2004, Bari et al., 2001) and females perspective, in separate studies (Becker et al., 1996, Bohman et al., 1992, Gates et al., 2004, Ghaffar et al., 2009, Khan et al., 2008a), yet the comparison of risk factors in both genders within the single study is lacking in literature. This current analysis provides an assessment of risk factors of HCV in both genders to ensure that all potential risk factors which could not be discovered in overall analysis have been identified in these separate analyses. It is worth mentioning that certain variables like barber shave, sharing of razors and nail cutting from barber, extramarital marital relationship and removal of unwanted hairs from barber shops are related to male subjects only. In contrast, some factors such as cesarean section, history of abortion/D&C and ear/nose piercing belong to female only. Therefore, a separate analysis of such risk factors followed by the general factors is a mandatory task to ascertain at present. It is pertinent to mention that interaction among area (urban/rural) and region (region/south) before stratifications was checked which showed significant effect (p=0.036).

Meanwhile, it is noticed that patients having either urban or rural settings certainly has different lifestyle, education level, occupational hazards and accessibility to health services. Similarly, environmental hazards, hospital facilities and level of awareness also vary considerably in both urban and the rural areas. Therefore, this inspired the researcher to develop independent models for male and female populations in view of both urban and rural settings of patients. Thus, male and female patients' data is further subdivided into urban-males, rural males; urban-females and rural-females to get further insight of potential risk factor of HCV at their very place.

Currently, just as the prior analysis ended up with overall sample, a similar univariate and multivariate logistic regression analysis is opted below to search for the possible risk factors of HCV among both genders. A gender-wise univariate logistic regression analysis accompanied by the percentage comparison of each socio-demographic variable in cases and controls is explained for both males and females independently in **Table 4.6**. Similarly, all other variables with their relevant analysis are explicated in **Table 4.7**. All variables having $p<0.20$ in univariate analysis are further selected for separate multivariate analysis to get final

selection of potential risk factors, their associated coefficients, odds ratio and 95% CI's. Thus, a multivariate logistic regression output is presented in Table 4.8 and Table 4.9 subsequently for males and females.

From **Table 4.7** results are compared within and between the gender-specific (male & female) models and area-specific (urban & rural) models. From **Table 4.8**, it is evident that only 9 factors emerged as potential risk factor of HCV in overall male model. While for overall female model in Table 4.9, there are 12 factors which established significant association with the disease outcome. The studies have shown that the females are more exposed to the infection because of greater exposures to gynecological examination, blood transfusion especially during pregnancy, surgery, delivery, hospitalization and ear/nose piercing etc (Ghaffar *et al.*, 2009).

On mutual comparison of gender-specific models, only five risk factors are found common in both genders. These included No. of persons sharing the room, patient's education level, family history of hepatitis C, blood transfusion and local migration/travelling. Besides these, history of blood donation, dental surgery, nails cuts and minor surgery by the barber are the additional risk factors in the males which enhanced the risk of disease by 2, 2, 3 and 2 folds respectively in patients who have positive history of these risk factors than others. Similarly, patient history of jaundice, ear/nose piercing, injected drugs, general surgery, history of abortion/D&C, Road crash injury and history of hospitalization are the additional pertinent risk factors of HCV in females with odds ratios 2, 4, 9, 2, 4, 3 and 2 folds, respectively. Among these risk factors, all have shown positive association with the outcome except patient's education which is negatively associated. Results indicate that patients' education can provide protection against this disease up to 84% in males and 90% in females with the higher level of education.

Similarly, while comparing these risk factors in urban/rural context, certain risk factors are different in both genders. It is also noticed that some risk factors which could not be identified in overall male and female models are additionally recognized in urban/rural settings of data. For example, tattooing, sharing of razors and removal of un-wanted hairs with the communal razors in "Hammams" are the additional risk factors of urban-males which augmented the risk of disease by 3, 4 and 2 folds, respectively. Such kinds of "Hammams" are mostly situated in urban communities near railway stations, bus stops and other public places like students/employees/laborers hostels/flats. These risk factors clearly reflect an inadequate life-style and behavioral characteristics of the patients. One more risk

factor that is “history of passed hostel life at any stage” is found in rural-male model because the persons usually shift to towns and occupied hostels or common flats for better education and livelihood.

It is also noticed that some factors such as No. of persons per room, patient’s low educational status, family history of hepatitis C and blood transfusion are also the common factor of each model. Other factor *i.e* accidental needle stick is also observed as common factor, from urban/rural male models which doubled the risk of disease with its positive history. It is found that accidental needle stick injury is more commonly reported among paramedics, tailors, and men/women who tailored at home and might have shared needles that are contaminated during the course of their work. The results suggested that about 40% of male-cases and 56% female-cases had reported history of accidental needle stick.

Interestingly, it is observed that some factors such as patient’s education, family history of hepatitis C, dental surgery, general surgery, minor surgery by barber and local migration /travelling are also the common factors in overall model as well as models for both genders. These factors are already discussed in details in previous subsection 4.3.5. Therefore, the discussion will only be restricted to recently identified risk factor of HCV after comparing both models.

In present study, the highest estimated risk is associated with injected drug use in females. This factor is also widely reported as a potential risk factor of HCV in literature. The current data reveal that the frequency of drug users is petty-low in both male and female cases *i.e* 8. 06% and 5. 94%, however, this factor appeared as a significant risk factor in overall female and urban-female models suggesting the corresponding risk of disease increases by 9 and 16 folds for the women who had history of injecting drug use, even once compared to others. A study from United States reported that the strongest risk factor of HCV infection is injecting drug use (IDU) which increases the risk by 149 folds (Armstrong *et al*. 2006). While a recent study in Baluchistan, Pakistan reveals similar finding but with different odds ratio 29. 95 (95%CI: 7. 06-127. 02) (Ahmed *et al.*, 2012).

The researcher further inquired and investigated that many of the females are receiving injected drugs from their spouses who had been drug users. Moreover, the patients belonging to Gujranwala, DG Khan, and Sargodha divisions exposed that there they are receiving injected drugs from the nearly situated drug stores (pharmacy) on payment. Luby *et al*.(1997) and Bari *et al*.(2001) carried case control studies from Hafizabad and Rawalpindi

districts of Punjab and found an insignificant association between illicit drug use and HCV disease. However, several other studies by Kaldor *et al* (1992), Darwish *et al* (1993), Mendes-Correa *et al* (2005) and He *et al* (2011) reported consistent findings and established the significant association with disease in Sydney, Egypt, Brazil and China, respectively.

Another important risk factor is observed in overall female, urban/rural-female models *i.e* "history of abortion/D&C" which increased the risk of disease by 4, 4 and 5 times, respectively. Overall about 37% female-cases reported the history of abortion/D&C. In a case control study, Ghaffar *et al*.(2009) could not establish an association of this factor with HCV among the women of reproductive age in Quetta, Pakistan. They found that about 24% and 17% of index women have or had the history of abortion and D&C, respectively. On the contrary, Habib *et al*.(2001) claimed that cesarean section or abortion (OR=1.4, 95% CI:1.0-1.9) is an evident factor of HCV Egyptian women. A comparable proportion of abortion is reported (21%) by Karaca *et al*. (2006) in Turkish population and suggested that "HCV may be transmitted during an abortion performed without sterilization". Other parallel studies by Merle *et al*.(1999) and Karmochkine *et al*.(2006a) from France also supported the recent study findings.

One more risk factor of HCV i.e. "history of hospitalization" is identified in female multivariate logistic regression model with odds ratio 1.6 (95% CI:1.03-2.6). This suggests that the risk of disease becomes double in patients who have previous history of hospitalization compared to others. The exact same odds ratio (OR=2.0) is identified in a case control study from France (Karmochkine et al., 2006a) and supported the findings of present study. Our data reveals that history of hospitalization is significantly higher in females (62%) compared to males (48%). Whilst this proportion of history of hospitalization is reported as (45%) in France population (Karmochkine et al., 2006a). Another study by Akhtar *et al*. (2004) reported the similar finding among the volunteer blood donors in Karachi, Pakistan. Some other studies from different regions of the world established the association of this risk factor with the disease in a parallel case control studies. For example, In a multivariate logistic regression; Chiaramonte *et al* (1996), Nguyen *et al* (2007) and Kandeel *et al*.(2012) identified significant relationship for this factor in Italy, Vietnam and Egypt, respectively. However, Mele *et al*.(1994) described a contrary finding with insignificant status of previous hospitalization with HCV infection.

Other factor like Blood transfusion is the common risk factor of HCV in Pakistan. In females, the researcher noticed two main factors, which lead to blood transfusion, are a

caesarean section or treatment for a road traffic-crash injury. In Pakistan, the caesarean section rate in a tertiary care hospital of Pakistan ranges from 17.8 to 31.2% (Eusaph *et al.*, 2011) and the average blood loss during caesarean is twice that of women undergoing vaginal births. Among Pakistani patients who have undergone a caesarean section, about 15% are transfused (Khan *et al.*, 2006), higher than a comparative study conducted in the USA (3.2%) (Rouse *et al.*, 2006). Blood used for transfusion in Pakistan is infrequently screened for infection, and the use of professional donors further increase the risk of contaminated blood (Shaheryar, 2012). Injury and hemorrhage, as a result of a road traffic crash, is likely to lead to transfusion, surgery and hospitalization and explain the observed association.

A socio-economic factor "No. of persons sharing the room" is included in both models as a continuous variable and showed positive association with the disease. In contrast, Luby *et al*.(1997) could not establish significant association for the family size and (Mean No. of persons per household: cases=5.5, controls= 6.7) as well as the persons sharing per room with the disease (Mean number of persons per room: cases=2.8, controls=2.1). However, the present study reveals that" number of persons sharing the room" is the potential socio-economic factor with comparable odds ratios in males and females 1.26 (95% CI: 1.16-1.37) and 1.19 (95% CI: 1.08-1.31), respectively. This suggests that by adding one more person per room the risk of disease increases by 1.26 and 1.19 folds in male and female patients, respectively. Abbas *et al*.(2008) found that about 32.5% of cases share the room with greater than 3 persons in Karachi, Sindh, Pakistan. The current data indicates that on the average 3 to 5 persons share the room in province, Punjab.

Other important factors such as ear/nose piercing, travelling are also reported as significant by Neal *et al* (1994) and He *et al* (2011) in UK and China, respectively. While on the other hand, Kim *et al* (1996), Luby *et al* (1997), Mishra *et al* (2003) in Korea, Pakistan, and to gather from the male and female veterans of the North Florida and South Georgia; these are reported as insignificant. Moreover, some more considerable risk factors in males are the History of blood donation, nails cut and minor surgery by barber which increase the risk of disease by 2, 3 and 2 folds, respectively.

It is pertinent to mention that interaction among area (urban/rural) and region (region/south) before stratifications was checked which showed significant effect (p=0.036).

Table 4.6: Gender-Wise Comparison of Socio-Demographic Variables

(A Univariate Analysis)

Variables	Category	Male				Female			
		Controls n=364	Cases n=397	Univariate logistic regression	p-value	Controls n=336	Cases n=303	Univariate logistic regression	p-value
		n (%)	n (%)	Odds ratio (95%CI)		n (%)	n (%)	Odds ratio (95%CI)	
Age group	<=34	215(59.07%)	198(49.87%)	1		176(52.38%)	130(42.90%)	1	
	>34	149(40.93%)	199(50.13%)	1.45(1.09-1.93)	0.011	160(47.62%)	173(57.10%)	1.46(1.07-2.00)	.017
Marital Status	Never married	92(25.27%)	65(16.37%)	1		105(31.25%)	34(11.22%)	1	
	Ever married	272(74.73%)	332(83.63%)	1.73(1.21-2.47)	0.003	231(68.75%)	269(88.78%)	3.60(2.35-5.50)	.000
Family status	Single	186(51.10%)	189(47.61%)	1		165(49.11%)	126(41.58%)	1	
	Nuclear	178(48.90%)	208(52.39%)	1.15(0.87-1.53)	**0.336**	171(50.89%)	177(58.42%)	1.36(0.99-1.85)	.057
Residential area	Urban	122(33.52%)	172(43.32%)	1		134(39.88%)	124(40.92%)	1	
	Rural	242(66.48%)	225(56.68%)	1.52(1.13-2.04)	0.006	202(60.12%)	179(59.08%)	1.04(0.76-1.43)	**.788**
House status	Owned	289(79.40%)	309(77.83%)	1		271(80.65%)	232(76.57%)	1	
	Rented	75(20.60%)	88(22.17%)	1.09(0.78-1.55)	**0.600**	65(19.35%)	71(23.43%)	1.28(0.87-1.86)	**.208**
Family size		7.06±3.39	7.67±3.99	1.046(1.01-1.09)	0.025	7.72 ± 4.01	7.34 ± 3.54	0.97(0.93-1.02)	**0.207**
No. of person per room		3.24±1.57	4.59 ± 3.19	1.28(1.19-1.38)	<0.0001	3.38±1.73	4.42±2.78	1.24(1.15-1.34)	<0.0001
Income group	<=5000	80(21.98%)	79(19.90%)	1	**0.530**	66(19.64%)	103(33.99%)	1	.000
	5000-15000	248(68.13%)	266(67.00%)	0.76(0.34-1.66)	0.484	222(66.07%)	170(56.11%)	0.49(0.34-0.71)	.000
	15000-25000	23(6.32%)	35(8.82%)	0.82(0.39-1.72)	0.601	24(7.14%)	18(5.94%)	0.48(0.24-0.95)	.036
	>25000	13(3.57%)	17(4.28%)	1.16(0.48-2.84)	0.739	24(7.14%)	12(3.96%)	0.32(0.15-0.68)	.003
Patient education	None	112(30.77%)	230(57.93%)	1	<0.0001	104(30.95%)	209(68.98%)	1	.000
	1-5	27(7.42%)	42(10.58%)	0.76(0.44-1.29)	<0.0001	63(18.75%)	36(11.88%)	0.28(0.18-0.46)	.000
	6-10	124(34.07%)	101(25.44%)	0.40(0.28-0.56)		80(23.81%)	46(15.18%)	0.29(0.19-0.44)	.000
	11-16	101(27.75%)	24(6.05%)	0.12(0.07-0.19)		89(26.49%)	12(3.96%)	0.07(0.04-0.13)	.000
Father's education	No	261(71.70%)	316(79.60%)	1		220(65.48%)	252(83.17%)	1	
	Yes	103(28.30%)	81(20.40%)	0.065(0.47-0.90)	0.011	116(34.52%)	51(16.83%)	0.38(0.26-0.56)	.000
Mother's education	No	288(79.12%)	330(83.12%)	1		258(76.79%)	258(85.15%)	1	
	Yes	76(20.88%)	67(16.88%)	0.77(0.53-1.11)	0.159	78(23.21%)	45(14.85%)	0.58(0.38-0.87)	.008

Table 4.7: Gender-Wise percentage comparison of risk factors and Univariate Analysis

Variables	Category	Male				Female			
		Controls n=364	Cases n=397	Univariate logistic regression	p-value	Controls n=336	Cases n=303	Univariate logistic regression	p-value
		n(%)	n(%)	Odds ratio (95%CI)		n(%)	n(%)	Odds ratio (95%CI)	
Family history of Hepatitis	No	316(86.81%)	273(68.77%)			278(82.74%)	207(68.32%)		
	Yes	48(13.19%)	124(31.23%)	3.0(2.07-4.33)	<0.0001	58(17.26%)	96(31.68%)	2.22(1.53-3.23)	<0.0001
Patient history of Jaundice	No	302(82.97%)	313(78.84%)			297(88.39%)	237(78.22%)		
	Yes	62(17.03%)	84(21.16%)	1.31(0.91-1.88)	0.149	39(11.61%)	66(21.78%)	2.12(1.38-3.26)	.001
Family history of jaundice	No	344(94.51%)	377(94.96%)			313(93.15%)	284(93.73%)		
	Yes	20(5.49%)	20(5.04%)	0.91(0.483-1.73)	0.778	23(6.85%)	19(6.27%)	0.91(0.49-1.71)	0.770
Patient history of liver disease	No	348(95.60%)	339(85.39%)			317(94.35%)	246(81.19%)		
	Yes	16(4.40%)	58(14.61%)	4.00(2.10-6.6)	<0.0001	19(5.65%)	57(18.81%)	3.87(2.24-6.67)	<0.0001
Family history of liver disease	No	354(97.25%)	384(96.73%)			316(94.05%)	286(94.39%)		
	Yes	10(2.75%)	13(3.27%)	1.20(0.52-2.77)	0.672	20(5.95%)	17(5.61%)		
History of blood transfusion	No	286(78.57%)	246(61.96%)			242(72.02%)	141(46.53%)		
	Yes	78(21.43%)	151(38.04%)	2.25(1.63-3.11)	<0.0001	94(27.98%)	162(53.47%)	2.96(2.13-4.11)	<0.0001
History of blood donation	No	274(75.27%)	246(61.96%)			305(90.77%)	279(92.08%)		.<0.0001
	Yes	90(24.73%)	151(38.04%)	1.87(1.37-2.56)	<0.0001	31(9.23%)	24(7.92%)		
Dental Surgery	No	279(76.65%)	223(56.17%)			252(75.00%)	156(51.49%)		
	Yes	85(23.35%)	174(43.83%)	2.56(1.87-3.50)	<0.0001	84(25.00%)	147(48.51%)	2.83(2.02-3.95)	<0.0001

Kidney dialysis	No	364(100.00%)	396(99.75%)	-	-	336(100.0%)	303(100.00%)	-	-
	Yes	0(0.00%)	1(0.25%)	-	-	-	-	-	-
Tattooing	No	324(89.01%)	324(81.61%)			318(94.64%)	248(81.85%)		
	Yes	40(10.99%)	73(18.39%)	1.83(1.21-2.76)	0.005	18(5.36%)	55(18.15%)	3.92(2.24-6.84)	<0.0001
Ear/Nose piercing	No	303(83.24%)	344(86.65%)			69(20.54%)	29(9.57%)		
	Yes	61(16.76%)	53(13.35%)			267(79.46%)	274(90.43%)	2.44(1.53-3.89)	<0.0001
More than one marriage	No	334(91.76%)	360(90.68%)			328(97.62%)	291(96.04%)		
	Yes	30(8.24%)	37(9.32%)	1.14(0.69-1.89)	0.600	8(2.38%)	12(3.96%)	1.69(0.68-4.19)	.257
Injected drugs, even once	No	353(96.98%)	365(91.94%)			328(97.62%)	285(94.06%)		
	Yes	11(3.02%)	32(8.06%)	2.81(1.40-5.67)	0.004	8(2.38%)	18(5.94%)	2.59(1.11-6.05)	.028
Major/Minor Surgery	No	299(82.14%)	269(67.76%)			256(76.19%)	151(49.83%)		
	Yes	65(17.86%)	128(32.24%)	2.19(1.56-3.08)	0.000	80(23.81%)	152(50.17%)	3.22(2.30-4.51)	<0.0001
Accidental needle stick	No	283(77.75%)	237(59.70%)			224(66.67%)	133(43.89%)		
	Yes	81(22.25%)	160(40.30%)	2.36(1.72-3.24)	0.000	112(33.33%)	170(56.11%)	2.56(1.85-3.52)	<0.0001
History of Angiography/Angioplasty	No	356(97.80%)	384(96.73%)			328(97.62%)	299(98.68%)		
	Yes	8(2.20%)	13(3.27%)	1.51(0.62-3.68)	0.368	8(2.38%)	4(1.32%)	1.82(0.54-6.12)	.331
History of endoscope or gastroscope	No	350(96.15%)	338(85.14%)			327(97.32%)	278(91.75%)		
	Yes	14(3.85%)	59(14.86%)	4.36(2.39-7.96)	0.000	9(2.68%)	25(8.25%)	3.27(1.50-7.12)	.003
History of Colonoscopy/Catheterization	No	340(93.41%)	371(93.45%)			313(93.15%)	268(88.45%)		
	Yes	24(6.59%)	26(6.55%)			23(6.85%)	35(11.55%)	1.78(1.02-3.08)	.041
Road crash injury	No	283(77.75%)	281(70.78%)			293(87.20%)	236(77.89%)		
	Yes	81(22.25%)	116(29.22%)	1.44(1.04-2.0)	0.029	43(12.80%)	67(22.11%)	1.93(1.27-2.94)	.002
History of cuts	No	237(65.11%)	183(46.10%)			209(62.20%)	121(39.93%)		
	Yes	127(34.89%)	214(53.90%)	2.18(1.63-2.92)	0.000	127(37.80%)	182(60.07%)	2.48(1.80-3.40)	<0.0001
Hospitalization	No	235(64.56%)	206(51.89%)			210(62.50%)	115(37.95%)		
	Yes	129(35.44%)	191(48.11%)	2.18(1.63-2.92)	0.000	126(37.50%)	188(62.05%)	2.72(1.98-3.75)	<0.0001
History of injections/Intravenous drips	No	151(41.48%)	66(16.62%)			129(38.39%)	75(24.75%)		
	Yes	213(58.52%)	331(83.38%)	4.00(2.54-4.98)	0.000	207(61.61%)	228(75.25%)	1.89(1.35-2.66)	<0.0001
History of branula insertion	No	227(62.36%)	170(42.82%)			203(60.42%)	116(38.28%)		
	Yes	137(37.64%)	227(57.18%)	2.21(1.65-2.96)	0.000	133(39.58%)	187(61.72%)	2.46(1.79-3.38)	<0.0001

Acupuncture	No	363(99.73%)	395(99.50%)			333(99.11%)	303(100.00%)		
	Yes	1(0.27%)	2(0.50%)	1.84(0.167-20.36)	0.620	3(0.89%)	0(0.00%)		
Ever imprisoned	No	339(93.13%)	355(89.42%)			324(96.43%)	292(96.37%)		
	Yes	25(6.87%)	42(10.58%)	1.60(0.96-2.69)	0.073	12(3.57%)	11(3.63%)	1.02(0.44-2.34)	.968
Local migration/Travelling	No	321(88.19%)	237(59.70%)			291(86.61%)	185(61.06%)		
	Yes	43(11.81%)	160(40.30%)	5.04(3.46-7.34)	0.000	45(13.39%)	118(38.94%)	4.12(2.79-6.09)	<0.0001
Foreign Travelling	No	357(98.08%)	373(93.95%)			325(96.73%)	301(99.34%)		
	Yes	7(1.92%)	24(6.05%)	3.28(1.40-7.71)	0.006	11(3.27%)	2(0.66%)		
Minor surgery by barber	No	343(94.23%)	298(75.06%)			316(94.05%)	274(90.43%)		
	Yes	21(5.77%)	99(24.94%)	5.50(3.31-8.91)	0.000	20(5.95%)	29(9.57%)	1.67(0.92-3.02)	.089
Eye surgery	No	353(96.98%)	375(94.46%)			324(96.43%)	283(93.40%)		
	Yes	11(3.02%)	22(5.54%)	1.88(0.90-3.94)	0.093	12(3.57%)	20(6.60%)	1.91(0.92-3.97)	.084
Ear wax removal from hospital	No	360(98.90%)	383(96.47%)			332(98.81%)	296(97.69%)		
	Yes	4(1.10%)	14(3.53%)	3.29(1.073-10.09)	0.037	4(1.19%)	7(2.31%)	1.96(0.57-6.77)	
Have you had passed hostel life?	No	283(77.75%)	270(68.01%)			310(92.26%)	269(88.78%)		
	Yes	81(22.25%)	127(31.99%)	1.64(1.19-2.27)	0.000	26(7.74%)	34(11.22%)	1.51(0.88-2.58)	.134
Sharing of nail cutter	No	128(35.16%)	83(20.91%)			121(36.01%)	61(20.13%)		
	Yes	236(64.84%)	314(79.09%)	2.052(1.48-2.84)	0.000	215(63.99%)	242(79.87%)	2.23(1.56-3.20)	<0.0001
Sharing of syringes	No	345(94.78%)	372(93.70%)			324(96.43%)	289(95.38%)		
	Yes	19(5.22%)	25(6.30%)	1.22(0.66-2.26)	0.525	12(3.57%)	14(4.62%)	1.31(0.60-2.87)	.504
Sharing of tooth brush/Miswak	No	304(83.52%)	292(73.55%)			270(80.36%)	224(73.93%)		
	Yes	60(16.48%)	105(26.45%)	1.82(1.28-2.60)	0.001	66(19.64%)	79(26.07%)	1.44(0.99-2.09)	.053
Sharing of razors(Male only)	Never	283(77.75%)	207(52.14%)			-	-	-	-
	Rarely	23(6.32%)	34(8.56%)	3.21(2.34-4.40)	0.000	-	-	-	-
	Often	58(15.93%)	156(39.29%)			-	-	-	-
Barber shave(Male only)	No	200(54.95%)	125(31.57%)			-	-	-	-
	Yes	164(45.05%)	271(68.43%)	2.64(1.97-3.55)	0.000	-	-	-	-
Under shave from barber shop(Male only)	No	247(67.86%)	149(37.82%)			-	-	-	-
	Yes	117(32.14%)	245(62.18%)	3.5(2.57-4.69)	0.000	-	-	-	-
Armpit shave from	No	262(71.98%)	236(59.75%)			-	-	-	-

barber(Male only)	Yes	102(28.02%)	159(40.25%)	1.73(1.28-2.35)	0.000	-	-	-	-
Extra-marital Marital relation (Male only)	No	206(56.75%)	175(44.08%)			-	-	-	-
	Yes	135(37.19%)	190(47.86%)	1.21(0.89-1.64)	0.230	-	-	-	-
	Non-response	22(6.06%)	32(8.06%)			-	-	-	-
Alcohol use(Male only)	No	357(98.08%)	385(96.98%)			-	-	-	-
	Yes	7(1.92%)	12(3.02%)	1.59(0.619-4.082)	0.335	-	-	-	-
Cesarean Section (Female only)	No	-	-	-	-	224(66.67%)	168(55.45%)		
	Yes	-	-	-	-	112(33.33%)	135(44.55%)	1.52(1.11-2.08)	.009
History of abortion/D&C (female only)	No	-	-	-	-	273(81.49%)	190(62.71%)		
	Yes	-	-	-	-	62(18.51%)	113(37.29%)	5.00(3.33-7.27)	<0.0001
Place of Cesarean Section (Female only)	None	-	-	-	-	206(61.31%)	147(48.51%)	1	<0.0001
	Private hospital	-	-	-	-	50(14.88%)	62(20.46%)	1.74(1.13-2.67)	.011
	Govt. hospital	-	-	-	-	78(23.21%)	71(23.43%)	1.28(0.89-1.87)	.215
	Midwife/LHV	-	-	-	-	2(0.60%)	23(7.59%)	16.12(3.7-69.1)	<0.0001
Did you use eye connectors, lens etc(Female only)	No	-	-	-	-	322(95.83%)	293(96.70%)		
	Yes	-	-	-	-	14(4.17%)	10(3.30%)	0.78(0.34-1.79)	.566

Table 4.8: Output from Multiple Logistic Regression Models for Male patients only

Variable		B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Overall Male Model								
No. of persons per room		0.233	0.042	30.433	0.000	1.263	1.162	1.372
Patient's education `	1-5	-0.252	0.318	0.630	0.427	0.777	0.417	1.449
	6-10	-0.933	0.209	19.916	0.000	0.393	0.261	0.592
	11-16	-1.812	0.296	37.458	0.000	0.163	0.091	0.292
Family history of Hepatitis C		1.054	0.226	21.8	0.000	2.87	1.843	4.467
Blood Transfusion		0.788	0.196	16.134	0.000	2.199	1.497	3.229
History of blood donation		0.577	0.201	8.276	0.004	1.781	1.202	2.638
Dental Surgery		0.432	0.196	4.873	0.027	1.54	1.05	2.26
Nails cut from barber		0.884	0.2	19.58	0.000	2.419	1.636	3.578
Local migration/travelling		1.117	0.234	22.721	0.000	3.057	1.931	4.839
Minor surgery by barber		0.73	0.303	5.791	0.016	2.074	1.145	3.758
Urban-Male Model								
No. of persons per room		0.353	0.074	22.68	0	1.423	1.231	1.646
Patient's education	1-5	-0.601	0.588	1.044	0.307	0.548	0.173	1.736
	6-10	-0.682	0.361	3.569	0.05	0.506	0.249	1.026
	11-16	-2.428	0.644	14.224	0	0.088	0.025	0.311
Blood donation		1.048	0.407	6.618	0.01	2.851	1.283	6.333
Dental surgery		0.867	0.35	6.135	0.013	2.38	1.198	4.728
Tattooing		1.097	0.568	3.726	0.054	2.994	0.983	9.116
History of needle click		0.726	0.343	4.487	0.034	2.067	1.056	4.045
Sharing of razors for shaving		1.36	0.426	10.174	0.001	3.894	1.689	8.979
Nails cutting from barber		1.753	0.4	19.161	0	5.77	2.632	12.646
Removal of un-wanted hairs from the "Hammam"/hostels or common flats		0.669	0.369	3.276	0.07	1.952	0.946	4.026
Rural-Male Model								
Patient's education	1-5	0.091	0.412	0.049	0.825	1.095	0.488	2.458
	6-10	-1.362	0.281	23.476	0	0.256	0.148	0.444
	11-16	-2.276	0.368	38.345	0	0.103	0.05	0.211
Family history of Hepatitis C		1.394	0.314	19.663	0	4.033	2.177	7.469
Blood transfusion		1.221	0.264	21.453	0	3.39	2.022	5.683
History of blood donation		0.626	0.255	6.009	0.014	1.87	1.134	3.083
Nails cut from barber		0.915	0.323	8.021	0.005	2.496	1.325	4.701
Minor surgery by barber		0.802	0.384	4.371	0.037	2.23	1.051	4.729
Have you had passed hostel life?		0.62	0.3	4.275	0.039	1.859	1.033	3.347

Table 4.9: Output from Multiple Logistic Regression Models for Female Patients

Variable		B	S.E.	Wald	p-value	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Overall Female Model								
No. of persons per room		0.174	0.049	12.67	0.000	1.19	1.081	1.309
Patient's education	1-5	-1.71	0.313	29.896	0.000	0.181	0.098	0.334
	6-10	-0.959	0.272	12.393	0.000	0.383	0.225	0.654
	11-16	-2.263	0.38	35.467	0.000	0.104	0.049	0.219
Family history of Hepatitis C		0.674	0.258	6.809	0.009	1.962	1.183	3.256
Patient history of Jaundice		0.688	0.292	5.569	0.018	1.99	1.124	3.525
Blood Transfusion		0.993	0.222	20.016	0.000	2.699	1.747	4.17
Ear/Nose piercing		1.284	0.342	14.094	0.000	3.611	1.847	7.058
Injected drugs, even once		2.143	0.586	13.38	0.000	8.521	2.703	26.857
General Surgery		0.798	0.247	10.452	0.001	2.221	1.369	3.603
History of abortion/D&C		1.345	0.253	28.217	0.000	3.837	2.336	6.302
Road crash injury		0.903	0.313	8.31	0.004	2.466	1.335	4.556
Hospitalization		0.488	0.236	4.267	0.039	1.63	1.025	2.591
Local migration/Travelling		1.026	0.257	15.957	0.000	2.791	1.687	4.619
Urban-Female Model								
No. of persons per room		.164	.084	3.773	.052	1.178	.999	1.390
Patient's education	1-5	-1.384	.495	7.807	.005	.251	.095	.661
	6-10	-1.491	.478	9.713	.002	.225	.088	.575
	11-16	-3.544	.836	17.977	.000	.029	.006	.149
Family history of Hepatitis C		1.766	.476	13.782	.000	5.848	2.302	14.859
Ear/Nose piercing		1.997	.484	17.042	.000	7.370	2.855	19.024
Injected drugs, even once		2.752	.962	8.193	.004	15.677	2.381	103.211
History of needle click		.708	.361	3.852	.050	2.031	1.001	4.120
History of abortion/D&C		1.391	.391	12.678	.000	4.019	1.869	8.643
History of branula insertion		1.175	.357	10.812	.001	3.238	1.607	6.522
Rural-Female Model								
No. of persons per room		.168	.058	8.353	.004	1.183	1.056	1.326
Patient's education `	1-5	-1.899	.389	23.856	.000	.150	.070	.321
	6-10	-1.042	.329	10.044	.002	.353	.185	.672
	11-16	-2.180	.457	22.729	.000	.113	.046	.277
Blood Transfusion		1.157	.270	18.350	.000	3.182	1.874	5.403
History of abortion/D&C		1.507	.327	21.300	.000	4.515	2.380	8.564
History of endoscope or gastro scope		2.059	.837	6.057	.014	7.841	1.521	40.422
Have you had passed hostel life?		1.218	.451	7.296	.007	3.382	1.397	8.187

4.4.1 Goodness of Fit of the Models

Similar to the overall model, different statistical tests are applied to assess the adequacy of the fitted models on gender and area-specific models. From **Table 4.10**, the results of Omnibus and Hosmer & Lemeshow (HL) tests suggested that each model is overall significant as p-values>0.05 simultaneously. Both models hold equally good fit of the sampled data. However, values of Nagelkerke's $R^2$and AUROC are comparatively improved in female LR models as compared to male LR models.

Table 4.10: Summary Statistics Showing Adequacy of the Fitted Models

Tests	Male Models			Female Models		
	Overall Male	**Urban-Male**	**Rural-Male**	**Overall Female**	**Urban-Female**	**Rural-Female**
Omnibus test	χ^2 =321.99 (p=0.000)	χ^2 =144.56 (p=0.000)	χ^2 = 91.43 (p=0.000)	χ^2 = 43.27 (p=0.000)	χ^2 = 44.42 (p=0.000)	χ^2 = 167.12 (p=0.000)
Hosmer and Lemeshow (HL)	χ^2 = 14.32 (p=.394)	χ^2 = 4.32, (p=.837)	χ^2 = 4.48 (p=.812)	χ^2 =7.687 (p=0.465)	χ^2 =11.41 (p=.180)	χ^2 =5.80 (p=0.671)
Cox and Snell R^2	32.6%	39.0%	33.7%	39.6%	43.0%	35.5%
Nagelkerke's R^2	43.6%	53.0%	45.0%	52.8%	57.2%	47.4%
AUROC	83.7%	80.2%	81.3%	87.8%	84.0%	85.4%

4.5 Region-specific Logistic Regression Models

For convenience, geographically Punjab province can be divided into two main regions *i.e.* North Punjab and South Punjab. North Punjab comprises of six important administrative Divisions namely Lahore, Sahiwal, Faisalabad, Gujranwala, Sargodha and Rawalpindi. Whereas, Southern Punjab has three allied Divisions *i.e.* Multan, Bahawalpur and DG Khan. It is observable that people between these regions of Punjab are variant in socio-demographic, life-style, behavioral characteristics and health facilities. Therefore, two separate logistic regression models are also run for these two regions by assimilating data of each concerned Divisional Headquarter Hospital followed by similar modeling building

procedure as explained earlier. In each region, models are built on overall region-wise data; male only and female settings, keeping in view their real importance. The details of cases and controls in each sub-division of the data in both regions of Punjab are given in **Table 4.11**. It illustrates that administrative divisions which belong to north-region of Province Punjab, total sample size is n=1000 including 500 cases and 500 controls. This sample size is reduced to 400 patients for Southern Punjab in such a way that cases and controls are 200 each. Similarly, other classification of cases and controls entrusted for males and females are also given and same classified data is used to develop different models. The results of multivariate logistic regression from North and Southern regions of Punjab including model coefficients, Wald test, p-values, odds ratio and their 95% CIs are given in **Table 4.13** & **Table 4.14**

Table 4.11: Region-Wise Distribution of Sample Size

	North-Region			**South-Region**		
	Overall	**Male**	**Female**	**Overall**	**Male**	**Female**
Cases	500	284	216	200	113	87
Controls	500	261	239	200	103	97
Total	1000	545	455	400	216	184

Table 4.12: Region and Gender-specific %age Classification of Risk Factors

Variables	Category	Regional Classification							
		South Punjab				North Punjab			
		Male		Female		Male		Female	
		Count	Column N %	Count	Column N %	Count	Column N %	Count	Column N %
Marital Status	Never married	43	19.91%	44	23.91%	114	20.92%	95	20.88%
	Ever married	173	80.09%	140	76.09%	431	79.08%	360	79.12%
Single/Joint	Single	111	51.39%	85	46.20%	264	48.44%	206	45.27%
	Joint	105	48.61%	99	53.80%	281	51.56%	249	54.73%
Urban/Rural	Urban	73	33.80%	68	36.96%	221	40.55%	190	41.76%
	Rural	143	66.20%	116	63.04%	324	59.45%	265	58.24%
House status	Owned	190	87.96%	158	85.87%	408	74.86%	345	75.82%
	Rented	26	12.04%	26	14.13%	137	25.14%	110	24.18%
Age group	<=34	135	62.50%	107	58.15%	278	51.01%	199	43.74%
	>34	81	37.50%	77	41.85%	267	48.99%	256	56.26%
Patient Education	None	98	45.37%	89	48.37%	244	44.77%	224	49.23%
	1-5	22	10.19%	25	13.59%	47	8.62%	74	16.26%
	6-10	56	25.93%	42	22.83%	169	31.01%	84	18.46%
	11-16	40	18.52%	28	15.22%	85	15.60%	73	16.04%
Income group	<=5000	39	18.06%	35	19.02%	120	22.02%	134	29.45%
	5000-15000	141	65.28%	118	64.13%	373	68.44%	274	60.22%
	15000-25000	24	11.11%	21	11.41%	34	6.24%	21	4.62%
	>25000	12	5.56%	10	5.43%	18	3.30%	26	5.71%
Father Education	No	160	74.07%	129	70.11%	417	76.51%	343	75.38%
	Yes	56	25.93%	55	29.89%	128	23.49%	112	24.62%
Mother Education	No	172	79.63%	146	79.35%	446	81.83%	370	81.32%
	Yes	44	20.37%	38	20.65%	99	18.17%	85	18.68%
Family history of Hepatitis C	No	156	72.22%	133	72.28%	433	79.45%	352	77.36%
	Yes	60	27.78%	51	27.72%	112	20.55%	103	22.64%
Patient history of Jaundice	No	178	82.41%	149	80.98%	437	80.18%	385	84.62%
	Yes	38	17.59%	35	19.02%	108	19.82%	70	15.38%
Family history of jaundice	No	204	94.44%	174	94.57%	517	94.86%	423	92.97%
	Yes	12	5.56%	10	5.43%	28	5.14%	32	7.03%
Patient history of liver disease	No	194	89.81%	159	86.41%	493	90.46%	404	88.79%
	Yes	22	10.19%	25	13.59%	52	9.54%	51	11.21%
Family history of liver disease	No	206	95.37%	172	93.48%	532	97.61%	430	94.51%
	Yes	10	4.63%	12	6.52%	13	2.39%	25	5.49%
History of blood transfusion	No	147	68.06%	107	58.15%	385	70.64%	276	60.66%
	Yes	69	31.94%	77	41.85%	160	29.36%	179	39.34%
History of blood	No	123	56.94%	172	93.48%	397	72.84%	412	90.55%

donation	Yes	93	43.06%	12	6.52%	148	27.16%	43	9.45%
Dental Surgery	No	139	64.35%	130	70.65%	363	66.61%	278	61.10%
	Yes	77	35.65%	54	29.35%	182	33.39%	177	38.90%
Tattooing	No	176	81.48%	146	79.35%	472	86.61%	420	92.31%
	Yes	40	18.52%	38	20.65%	73	13.39%	35	7.69%
Ear/Nose piercing	No	179	82.87%	27	14.67%	468	85.87%	71	15.60%
	Yes	37	17.13%	157	85.33%	77	14.13%	384	84.40%
More than one marriage	No	183	84.72%	177	96.20%	511	93.76%	442	97.14%
	Yes	33	15.28%	7	3.80%	34	6.24%	13	2.86%
Injected drugs, even once	No	202	93.52%	178	96.74%	516	94.68%	435	95.60%
	Yes	14	6.48%	6	3.26%	29	5.32%	20	4.40%
History of needle click	No	154	71.30%	107	58.15%	366	67.16%	250	54.95%
	Yes	62	28.70%	77	41.85%	179	32.84%	205	45.05%
Sharing of syringes	No	201	93.06%	179	97.28%	516	94.68%	434	95.38%
	Yes	15	6.94%	5	2.72%	29	5.32%	21	4.62%
Sharing of tooth brush/Miswak	No	156	72.22%	136	73.91%	440	80.73%	358	78.68%
	Yes	60	27.78%	48	26.09%	105	19.27%	97	21.32%
Sharing of razors	No	112	51.85%	0	0.00%	378	69.36%	0	0.00%
	Yes	104	48.15%	0	0.00%	167	30.64%	0	0.00%
Barber shave	No	65	30.09%	0	0.00%	254	46.61%	0	0.00%
	Yes	151	69.91%	0	0.00%	291	53.39%	0	0.00%
Sharing of nail cutter	No	45	20.83%	43	23.37%	166	30.46%	139	30.55%
	Yes	171	79.17%	141	76.63%	379	69.54%	316	69.45%
Barber nails cutter	No	108	50.00%	0	0.00%	381	69.91%	0	0.00%
	Yes	108	50.00%	0	0.00%	164	30.09%	0	0.00%
Under shave from barber shop	No	94	43.52%	0	0.00%	302	55.41%	0	0.00%
	Yes	122	56.48%	0	0.00%	243	44.59%	0	0.00%
Armpit from barber	No	97	44.91%	0	0.00%	401	73.85%	0	0.00%
	Yes	119	55.09%	0	0.00%	142	26.15%	0	0.00%
Marital relation other than legal partner	No	65	30.09%	0	0.00%	316	58.09%	0	0.00%
	Yes	121	56.02%	0	0.00%	204	37.50%	0	0.00%
	Non-response	30	13.89%	0	0.00%	24	4.41%	0	0.00%
Caesarian section(if female)	No	0	0.00%	90	48.91%	0	0.00%	263	57.80%
	Yes	0	0.00%	94	51.09%	0	0.00%	192	42.20%
History of abortion (female only)	No	0	0.00%	130	70.65%	0	0.00%	336	73.85%
	Yes	0	0.00%	54	29.35%	0	0.00%	119	26.15%
History of Angiography/Angioplasty	No	212	98.15%	181	98.37%	528	96.88%	446	98.02%
	Yes	4	1.85%	3	1.63%	17	3.12%	9	1.98%
History of endoscope or gastroscopy	No	181	83.80%	175	95.11%	507	93.03%	430	94.51%
	Yes	35	16.20%	9	4.89%	38	6.97%	25	5.49%
History of Colonoscopy	No	207	95.83%	162	88.04%	504	92.48%	419	92.09%
	Yes	9	4.17%	22	11.96%	41	7.52%	36	7.91%
Road traffic accident	No	158	73.15%	154	83.70%	406	74.50%	375	82.42%
	Yes	58	26.85%	30	16.30%	139	25.50%	80	17.58%

History of cuts	No	115	53.24%	91	49.46%	305	55.96%	239	52.53%
	Yes	101	46.76%	93	50.54%	240	44.04%	216	47.47%
Hospitalization	No	106	49.07%	80	43.48%	335	61.47%	245	53.85%
	Yes	110	50.93%	104	56.52%	210	38.53%	210	46.15%
History of injections/Intravenous drips	No	57	26.39%	58	31.52%	160	29.36%	146	32.09%
	Yes	159	73.61%	126	68.48%	385	70.64%	309	67.91%
History of branula injection	No	93	43.06%	86	46.74%	304	55.78%	233	51.21%
	Yes	123	56.94%	98	53.26%	241	44.22%	222	48.79%
Acupuncture	No	214	99.07%	183	99.46%	544	99.82%	453	99.56%
	Yes	2	0.93%	1	0.54%	1	0.18%	2	0.44%
Ever imprisoned	No	201	93.06%	183	99.46%	493	90.46%	433	95.16%
	Yes	15	6.94%	1	0.54%	52	9.54%	22	4.84%
Local migration/Travelling	No	168	77.78%	147	79.89%	390	71.56%	329	72.31%
	Yes	48	22.22%	37	20.11%	155	28.44%	126	27.69%
Minor surgery by barber	No	176	81.48%	164	89.13%	465	85.32%	426	93.63%
	Yes	40	18.52%	20	10.87%	80	14.68%	29	6.37%
Eye surgery, stitching	No	201	93.06%	171	92.93%	527	96.70%	436	95.82%
	Yes	15	6.94%	13	7.07%	18	3.30%	19	4.18%
Ear wax removal from hospital	No	208	96.30%	183	99.46%	535	98.17%	445	97.80%
	Yes	8	3.70%	1	0.54%	10	1.83%	10	2.20%
Have you had passed hostel life?	No	151	69.91%	161	87.50%	402	73.76%	418	91.87%
	Yes	65	30.09%	23	12.50%	143	26.24%	37	8.13%

Table 4.13: Multiple Logistic Regression Model Output for North Punjab Region

Variables		B	S.E.	Wald	p-value	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Overall North Punjab Model								
Gender (Male)		.549	.166	11.009	.001	1.732	1.252	2.396
No. of persons per room		.203	.034	36.621	.000	1.225	1.147	1.308
Patient's education	1-5	-.846	.241	12.362	.000	.429	.268	.688
	6-10	-1.018	.192	28.265	.000	.361	.248	.526
	11-16	-1.728	.266	42.123	.000	.178	.105	.299
Family history of Hepatitis C		.842	.197	18.224	.000	2.322	1.577	3.418
Patient history of Liver disease		.719	.281	6.542	.011	2.053	1.183	3.563
Blood Transfusion		.933	.170	30.116	.000	2.541	1.821	3.546
Dental surgery		.502	.167	8.974	.003	1.652	1.189	2.293
Injected drugs, even once		.814	.387	4.429	.035	2.257	1.058	4.818
History of needle click		.424	.170	6.222	.013	1.528	1.095	2.132
Major/Minor Surgery		.740	.179	17.133	.000	2.097	1.477	2.977
History of injections/Intravenous drips		.376	.184	4.155	.042	1.456	1.015	2.090
History of branula insertion		.507	.171	8.854	.003	1.661	1.189	2.320
Minor surgery by barber		.612	.275	4.944	.026	1.844	1.075	3.161
Male only								
No. of persons per room		.231	.047	24.574	.000	1.260	1.150	1.381
Patient's education	1-5	.264	.401	.433	.511	1.302	.593	2.855
	6-10	-.778	.243	10.289	.001	.459	.285	.739
	11-16	-1.407	.347	16.437	.000	.245	.124	.484
Family history of Hepatitis C		1.183	.284	17.349	.000	3.266	1.871	5.700
Blood Transfusion		.712	.236	9.108	.003	2.038	1.283	3.235
Blood Donation		.502	.243	4.267	.039	1.652	1.026	2.661
Dental Surgery		.451	.230	3.839	.050	1.570	1.000	2.464
Sharing of Razors		.866	.241	12.935	.000	2.378	1.483	3.812
Major/Minor Surgery		.701	.252	7.722	.005	2.016	1.230	3.306
History of injections/Intravenous drips		.775	.249	9.726	.002	2.171	1.334	3.534
Minor surgery by barber		.795	.335	5.610	.018	2.214	1.147	4.272
Female Only								
No. of persons per room		.200	.061	10.797	.001	1.221	1.084	1.375
Patient's education	1-5	-2.110	.387	29.777	.000	.121	.057	.259
	6-10	-1.536	.377	16.563	.000	.215	.103	.451
	11-16	-2.245	.482	21.704	.000	.106	.041	.272
Family history of Hepatitis C		1.078	.345	9.780	.002	2.939	1.495	5.776
Patient past history of Jaundice		.789	.399	3.912	.048	2.202	1.007	4.815
Blood Transfusion		1.360	.294	21.365	.000	3.898	2.189	6.939
Ear/Nose piercing		2.030	.452	20.169	.000	7.617	3.140	18.475
Sharing of nail cutter		.789	.335	5.555	.018	2.201	1.142	4.242
Major/Minor Surgery		.734	.305	5.789	.016	2.083	1.146	3.788
History of abortion/D&C		1.646	.341	23.250	.000	5.185	2.656	10.122
Road crash injury		1.186	.402	8.697	.003	3.273	1.488	7.196
History of branula insertion		1.011	.292	11.995	.001	2.750	1.551	4.874

Table 4.14: Multiple Logistic Regression Model Output for South Punjab Region

Variables		B	S.E.	Wald	p-value	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Overall South Punjab Model								
Patient's education	1-5	-0.892	0.378	5.571	0.018	0.41	0.195	0.86
	6-10	-0.669	0.292	5.242	0.022	0.512	0.289	0.908
	11-16	-2.656	0.445	35.69	0	0.07	0.029	0.168
Patient history of liver disease		1.127	0.419	7.247	0.007	3.087	1.359	7.013
Blood Transfusion		0.63	0.253	6.188	0.013	1.877	1.143	3.082
History of blood donation		0.737	0.285	6.702	0.010	2.089	1.196	3.65
Tattooing		1.256	0.341	13.522	0.000	3.51	1.798	6.856
Sharing of tooth brush/Miswak		0.569	0.27	4.428	0.035	1.766	1.04	2.998
History of endoscopy or gastroscope		2.625	0.644	16.613	0.000	13.798	3.906	48.742
Male only								
Patient's education	1-5	-1.914	0.657	8.488	0.004	0.147	0.041	0.535
	6-10	-1.414	0.509	7.732	0.005	0.243	0.09	0.659
	11-16	-2.689	0.601	19.994	0.000	0.068	0.021	0.221
Patient history of liver disease		2.235	0.916	5.956	0.015	9.349	1.553	56.288
Blood Transfusion		1.031	0.419	6.067	0.014	2.805	1.235	6.373
Nails cutting from barber		0.981	0.521	3.548	0.060	2.667	0.961	7.402
Removal of un-wanted hairs from the "Hammam"/hostels or common flats		1.608	0.536	9.019	0.003	4.995	1.748	14.271
History of endoscope or gastroscope		2.146	0.962	4.976	0.026	8.551	1.298	56.355
Minor surgery by barber		1.539	0.737	4.361	0.037	4.658	1.099	19.738
Fe-male only								
Family status (Nuclear family)		-1.103	0.392	7.943	0.005	0.332	0.154	0.715
Patient's education	1-5	-0.791	0.518	2.327	0.127	0.454	0.164	1.253
	6-10	-0.324	0.441	0.538	0.463	0.724	0.305	1.718
	11-16	-2.4	0.725	10.95	0.001	0.091	0.022	0.376
Family history of Hepatitis C		1.331	0.458	8.459	0.004	3.785	1.543	9.28
History of accidental needle		0.695	0.375	3.431	0.064	2.004	0.96	4.182
History of abortion/D&C		1.653	0.432	14.618	0.000	5.22	2.238	12.178

Table 4.15: Summary Statistics Showing Adequacy of Fitted Models for North/South Regions of Punjab.

Tests	North-Punjab Models			South-Punjab Models		
	Overall	**Male only**	**Female only**	**Overall**	**Male**	**Female**
Omnibus test	χ^2 =388.3 (p=0.000)	χ^2 =208.9 (p=0.000)	χ^2 =283.2 (p=0.000)	χ^2 =144.9 (p=0.000)	χ^2 =135.2 (p=0.000)	χ^2 =71.2 (p=0.000)
Hosmer and Lemeshow (HL)	χ^2 =8.242 (p=0.410)	χ^2 =3.937 (p=.863)	χ^2 =8.781 (p=0.361)	χ^2 =7.69 (p=0.080)	χ^2 =9.14 (p=0.330)	χ^2 =5.69 (p=0.681)
Cox and Snell R^2	32.2%	31.9%	46.3%	39.6%	46.5%	32.1%
Nagelkerke' R^2	42.9%	42.5%	61.8%	52.8%	90.8%	42.8%
AUROC	83.9%	82.5%	90.9%	87.8%	82.7%	83.5%

4.5.1 Goodness of Fit of the Models

It represents a comparison of adequacy of the fitted models in both north and southern regions of the Province Punjab using different statistical tools. Omnibus and HL tests showed that all models are overall significant at p-value<0.05 and p-value>0.05 respectively. Highest values of Nagelkerke's $R^2$and AUROC are observed in South-male and North-female models *i.e* 90.8% and 90.9% respectively indicating very good fit. On the whole, it is evident that models for both regions of Province Punjab are adequately fitted.

4.5.2 Comparison and Interpretation of Region-specific Models

To understand the broaden picture of risk factors of hepatitis C in province Punjab, a comparison of north and southern regions of Punjab could be fairly beneficial to investigate relevant risk factors at their best place. This analysis would certainly enable us to suggest region wise preventive strategy to prevent people of the Province from the hazardous disease by understanding of allied risk factors in the regions. From **Table 4.13** & **Table 4.14** it is observed that 12 risk factor are found significant in overall north model whereas only seven in overall South model. Some risk factor such as patient's education, patient history of liver disease and blood transfusion are the common in both overall northern and southern models. Only the patient's education is negatively associated while all others have its positive association with the disease. In addition, other factors including family history of hepatitis C, dental surgery, accidental needle stick, major/minor surgery, History of injections and

branula insertion, cut on wounds by barber are the part of overall north-model. The estimated odds ratio of these significant risk factors suggested that with the positive history of every individual factor, the risk of disease increases by 2 times. Similarly, among the socio-demographic factors only one factor i.e No. of persons the room is the significant factor which raises the risk of disease by 1.225 times. Importantly, it is noticed that only patient's education may reduce the risk of disease by 57% to 82%. The data reflected that about 45%-49% of the patients have no formal education in the regions. Compared to North region, some factors such as tattooing, sharing of toothbrush/Miswak and endoscope are the different risk factors in South Punjab having odds ratio 3.51, 1.76 and 13.80 respectively. Maximum likelihood of disease is associated with history of endoscopy which indicates inadequate, un-hygienic control in hospitals in Southern Punjab. The results also revealed the fact that sharing of tooth brush/Miswak is also an essential risk factor in this region along with odds ratio (OR=1.77). About 28% males and 26% females share their tooth brushes in Southern Punjab which reflect their unpleasant behavioral characteristic and insufficient awareness. Majority of the women had used tooth brush/Miswak with their partners at certain occasions whereas a number of the individuals have or had used Miswak placed at ablution place in Mosques. The results from male models are compared in both regions and found that some factors such as patient's education, blood transfusion and cut on wounds by barber are the common factors but with different odds ratios. Other factors including family H/O hepatitis C, blood donation, sharing of razors, surgical operations, and history of injections are the key factors of northern males. On contrary patients H/O liver disease (OR=9.349), nails cutting by barber (OR=2.67), removal of unwanted hairs from barber shops (OR=4.990) and endoscopy (OR= 4.658) are the essential factors of Southern males.

Similarly, on comparing the results from female models, only 3 factors *i.e* patient's education, family H/O hepatitis C and H/O abortion/D&C are comparable in South as well as Northern female patients. Some additional risk factors such as Patient past history of Jaundice, blood transfusion, Major/Minor Surgery, Road crash injury and history of branula insertion are the significant factors in north female patients which relates to the medical history. Whereas Ear/Nose piercing and sharing of nail cutter behavioral characteristics of the patients are found significant in north female patients along with their corresponding odds ratios 3.898, 7.617, 2.201, 2.083, 3.273 and 2.750 respectively. 95% CIs of each odds ratio are also given in the attached columns. This comparison indicates that some risk factors are different by gender, residential area and region of the Punjab province.

Chapter 5

Application of Artificial Neural Networks & Classification Trees Models

Keeping in view the other objectives of this study, in this chapter, application/usefulness of Artificial Neural Networks (ANN) and Classification Trees models have been introduced to analyze risk factors of hepatitis C in addition to Logistic regression model (Gold Standard). The results from overall data are solely compared and contrasted with Logistic regression. For statistical analysis, firstly, univariate logistic regression is performed for each variable and shortlisted predictors which are associated with the outcome at $p<0.20$ (Hosmer and Lemeshow, 2000). Then, these predictors are further included in modeling multivariate logistic regression and ANN models respectively, to identify most pertinent risk factors of HCV infection. All p-values are 2-tailed. The discriminatory power of each model to differentiate between cases and controls is assessed by use of the area under the Receiver Operating Characteristic (AUROC) curve. Moreover, before comparing the results from 3 approaches, it was ascertain that a right logistic regression selected.

5.1 Development of ANN Model

In this study, a popular Multi-layer feed-forward neural network is implied (Zurada and Lonial, 2011, Mohamed et al., 2011); where, risk factors are taken as the inputs and the presence or absence of HCV seropositivity as the output.

The input variables consisted of **factors** and **covariates;** in which factors are categorical independent variables (nominal or ordinal), while covariates are quantitative continuous variables (ages, family size, income etc). The covariates are rescaled initially by standardized rescaling to improve the performance of the ANN model (Tufféry, 2011).

Before training neural network, the whole data is randomly divided into two partitions *i.e* Training (80%) and Testing (20%) in such a way that pre-assigned cases to controls ratio does not affect. The training set should be a relatively larger portion of records as compared to testing set which is used to train the neural network. On completion of training, the simulated results then are further verified, counter checked on employing with testing set; an independent sample from the same data set. Testing data help to track errors and prevent from over-fitting during training process**.** As, it has been observed that ANN model exhibits a

drawback of over-fitting and this can be compensated by performing the partitioning of data, as proposed above. It is pertinent to mention that different ANN models were performed based on multiple divisions of data sets. During examination it was observed that portioning of data set (80% training 20% testing) was the best choice as it produces least sum of square of residuals. Moreover, different studies have also validated this proposed division.

'Batch method with a Scaled Conjugate Gradient' was applied as a useful optimization algorithm which reduces the error for the small data set (Tufféry, 2011). As our dependent variable is binary, (presence or absence) of HCV seropositivity, a sigmoid or logistic activation function is used at hidden nodes and at the output variable. The signals are processed through these activation functions transforming to the output node.

Different successive ANN models are developed to reach a best possible model but only one having overall good model performance is reported and with minimum value of sum of square of errors (SSE) (Maimon and Rokach, 2005). During model building, on every repeated trial the other parameters such as ROC curve are performed for examining network performance. The main architecture of the neural networks could be denoted as 33-12-1, means that 33 predictors, 12 hidden nodes and 1 output (target) variable.

Figure 5.1 gives illustration for the final output from the ANN fitted model, obtained by adopting the same procedure as described in the methods section. The rank and sequence of each factor are also illustrated in

Figure **5.1.** This is obtained by plotting normalized importance (by ANN) of each risk factor. According to the importance, 11 most important risk factors are finally obtained. Just like the LR finding, patients' education also has the top rank in the ANN model. Cut point at the 11^{th} number risk factor is considered because almost no change in normalized importance observed between the10^{th} and 11^{th}, while a sudden decrease occurred when the 12^{th} variable is included in the model. If the risk factors at rank no.12^{th} and 13^{th} be considered for selection, normalized importance becomes too least. It is therefore, decided that first 11 no. Factors be treated as most important. This selection of variable becomes more reliable when results of ANN are compared with LR model i.e. about 11 risk factors are significantly associated with disease in logistic regression analysis **Table 4.5** and with almost the same number of variables selected by ANN model.

5.1.1 Sensitivity Analysis

ANN models have been criticized for being “black box” due to their inability to generate interpretable coefficients (weights) for each predictor variable (Wang *et al.*, 2010). To compensate, sensitivity analysis is carried out, which computed the importance of each predictor in determining the neural network (Wang *et al.*, 2010). This is used to determine the relative importance of each input parameter in the model and help to rank the importance of different variables (Shi *et al.*, 2012). The importance of an independent variable is actually a measure of just how much the network’s model-predicted value varies for different values of the predictor variables. Moreover, the normalized importance is simply the ratio of importance values which are estimated by the final ANN model and the largest importance values, then expressing as percentages. The discriminatory power of each model to differentiate between cases and controls is assessed by use of the area under the receiver operating characteristic (AUROC) curve. Further, sensitivity and specificity analysis are preformed using OpenEpi Version 2.3.1 software for each model.

5.1.2 Model Performance

Overall performance for each model is assessed by area under the ROC (AUROC) curve, sensitivity and specificity analysis (at a cut point of 0.5 probability). For the LR model these measures are found as 83.8, 76.1, and 74.8%, respectively. Moreover, overall model classification is 75.4% correct. On the other hand, the overall correct classification for ANN model in training and testing sets of data is 77.3 and 79.8%, respectively. The area under the ROC curve is 84.7% in ANN model combining both training and testing data sets. Further, sensitivity and specificity of ANN model is 77.6, and 77% respectively, in training data while for testing data these parameters are 78.5 and 81.2%, respectively.

These results show that the model discriminating power is similar, comparing the two methods, from the ROC analysis, whilst sensitivity and specificity are higher in ANN models. This improvement in sensitivity and specificity in ANN may have been due to the inclusion of interaction effects in the model.

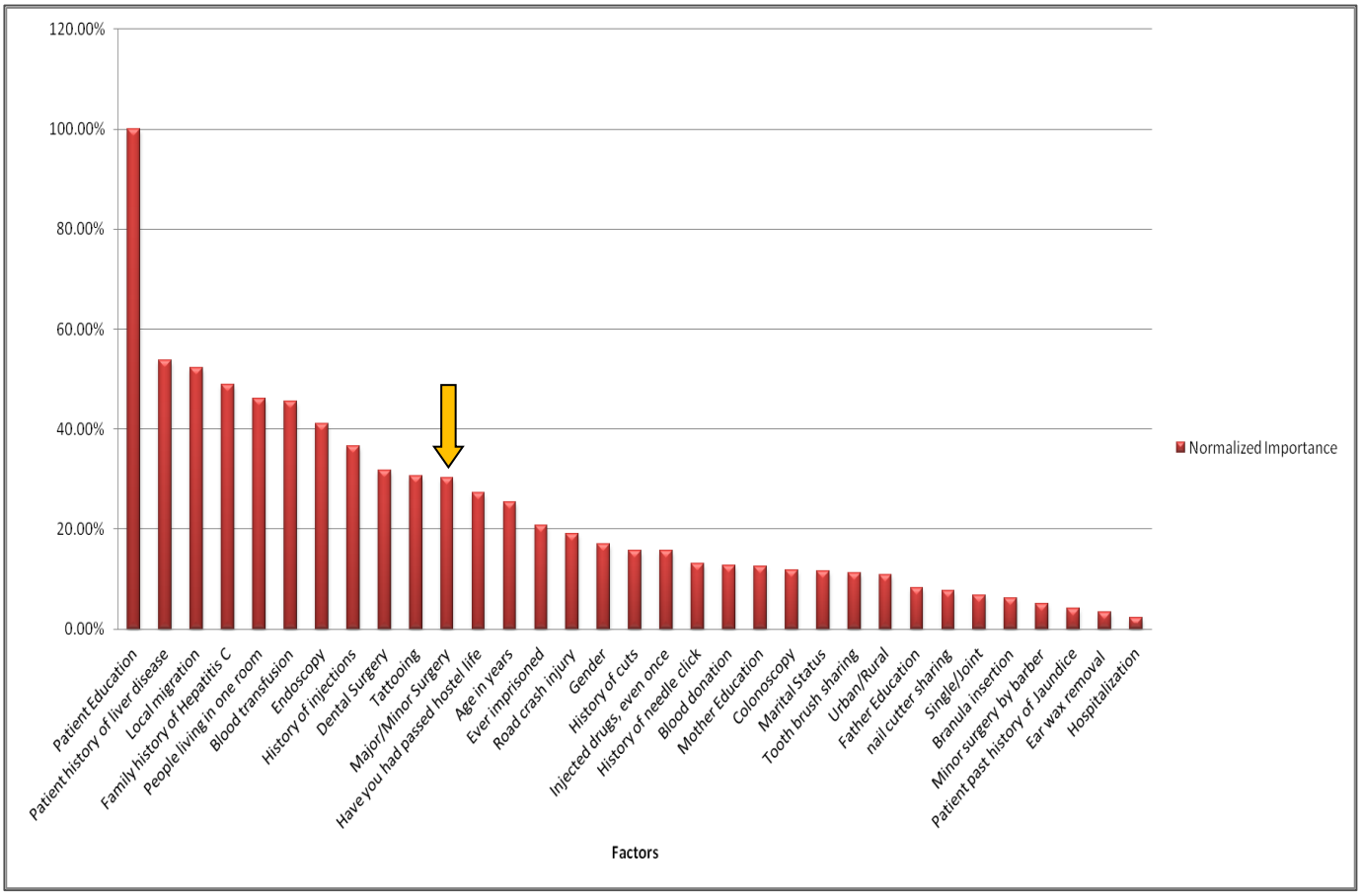


Figure 5.1: Independent Variable Importance Chart from ANN Model

Figure 5.1 shows the ranking of each input variable through the importance and normalized importance, obtained by ANN model. The importance of an independent variable is actually a measure of just how much the network's model-predicted value varies for different values of the predictor variables. Moreover, the normalized importance is simply the ratio of importance values which are estimated by the final ANN model and the largest importance values, then expressing as percentages.

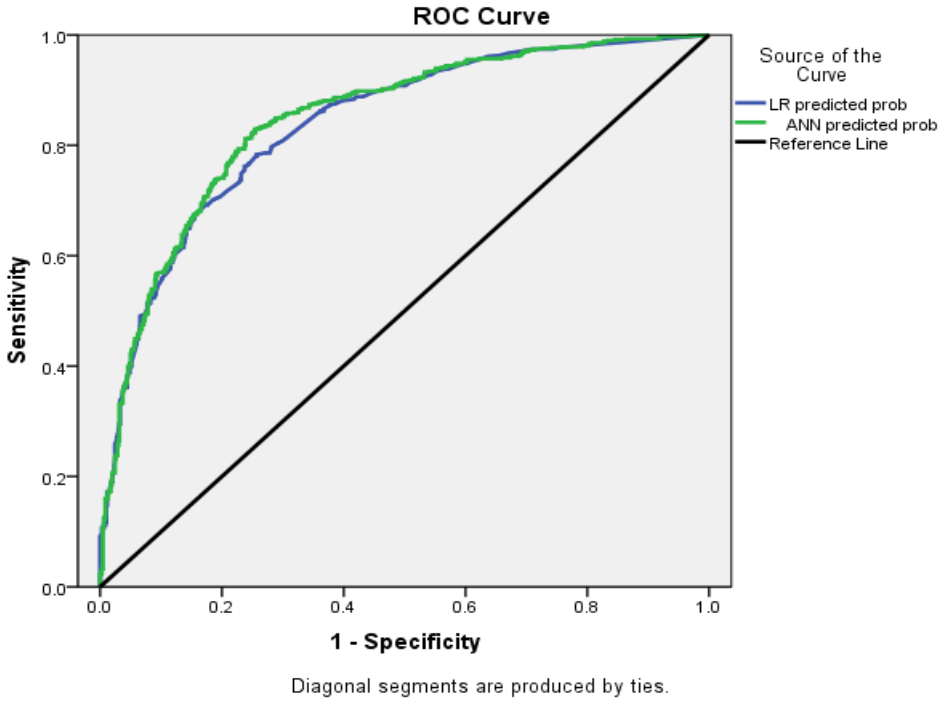


Figure 4.8: Comparison of ROC Curves in ANN and LR models

5.2 Comparison of Logistic Regression with ANN model

Interestingly, about 11 risk factors are significantly associated with the disease by logistic regression analysis **Table 4.5** and with almost the same number of variables selected by ANN model.

Figure 5.1. However, one risk factor from each model differed, "minor surgery by barber" is significant in LR but not in the ANN model. Similarly, "Number of persons sharing one room" is an important variable in ANN but not in LR. Thus, after combining the results from both models, we found 12 risk factors that are associated with disease status in Punjab province, Pakistan. These risk factors are: patient education, patient history of liver disease, local migration, family history of hepatitis C, number of persons sharing one room, history of blood transfusion, endoscopy, history of injections, dental surgery, tattooing, general surgery, and minor surgery by barber.

5.3 Discussion Based on Results from LR and ANN models

The purpose of the present study is to apply LR and ANN models to model and analyze the risk factors of HCV infection, but not to prove superiority of a neural network with that of conventional LR statistical modelling. The main findings suggested that use of an

ANN model is helpful for both variable selection and ranking of important variables in sequence of importance and the conclusion is supported by another study by (Sarle, 2000).

The strength of this study is the application of both models to real data, exploring the features of the data shown by each method. This helps us to account for the geographic, socio-demographic and personal living pattern of life in the study. On comparing the results from LR and ANN models 12 independent variables contributing towards HCV infection in the province (Punjab), Pakistan are identified. In a recent literature, it is found that among these factors some traditional risk factors such as history of blood transfusion, history of injections, dental surgery and general surgery are identified in other studies, carried out in other regions and settings (Souto et al., 2012, Ho et al., 2012b, Azevedo et al., 2012, He et al., 2011, Qureshi et al., 2009). Other factors include family history of hepatitis C (Ghias and Pervaiz, 2009c), patient history of liver disease, endoscopy (Karmochkine et al., 2006a), and tattooing (Bari et al., 2001, Ghias and Pervaiz, 2009c); which are reported but not so common. Low patient education (Shazi and Abbas, 2006), local migration, number of persons sharing one room, and minor surgery by barber are the other factors. This study documented that about 31.4% of cases reported a positive family history of hepatitis C. This augment the risk as 55% cases live in a joint family system and reported receiving no formal education (62.7%). Moreover, on the average, about 5 persons are found to share one room in overall data.

Our ANN model assigned the highest rank to a patient's education is the main factor which can play the pivotal role to reduce the burden up to 20%. For the LR method, odds ratios for each predictor variable can easily be calculated. But, in ANN models it is not easy to derive an odds ratio. However, one study has reported a method to compute an effective odds ratio for the ANN and compared the results with LR model bagging, reporting almost similar results (Green *et al.*, 2006)

There are many studies in which multivariable logistic regression is applied to find out the significant risk factors for HCV infection. For instance, a recent cross sectional study is conducted in Baluchistan, Pakistan, finding that HCV prevalence in this region is about 5.5%. After multivariable logistic regression, age>75 years, use of injected drugs, and being a healthcare worker are most strongly associated with HCV infection (Ahmed *et al.*, 2012). Another study, conducted from a peri-urban community of Karachi, identified the prevalence and risk factors of HCV infection in men and women separately. The logistic regression model identified that history of blood transfusion; injections and dental surgery are the main

risk factors in women. While extramarital relation, barber shave and increasing age showed the strongest association in men (Janjua *et al.*, 2010). Another study which used logistic regression analysis to explore risk factors of hepatitis C in males found, similarly, that blood transfusion and history of undergoing dental surgery are the main factors associated with disease status. A case control study (Khan et al., 2008a) among pregnant women at Karachi hospital, which used multivariate logistic regression ultimately identified three major risk factors associated with HCV infection; health care injections, hospitalizations and number of pregnancies. Similarly, many other researchers have applied the same logistic technique to evaluate risk factors for hepatitis C infection. In short, many studies have used this model for risk factors identification, but some shortcomings have been identified in their methods. For example, majority of them did not apply any diagnostic checking (outliers, influential values), multicollinearity prior to developing the final logistic regression model (Karmochkine et al., 2006b, Karim et al., 2011, Janjua et al., 2010). This study covers these limitations and also used ANN models as an alternative to logistic regression model to identify the effective risk factors of hepatitis C infection in Punjab, Pakistan. The ANN model is adequately built on overall data as indicated by AUROC, sensitivity and specificity of model.

The two methods have been compared in studying other diseases; for example, in diabetes, Salmonella typhimurium infections, and coronary artery bypass (Qin et al., 2005, Voss et al., 2002, Gao et al., 2004). Another study of the risk factors of uterine myomas using ANN model and LR demonstrated similar leading risk factors and on comparison with conventional statistics, ANN perform better, on measures of discrimination and provided powerful alternative to risk factors analysis (wie, 2007). The same conclusion is supported by other studies (Flaherty and Patterson, 2003, Dussol *et al.*, 2007) that on comparison ANN model is not more efficient than an LR model in discriminating the different parameters. Hart and Hart & Wyatt (Hart and Wyatt, 1990) presented a critique that "black box" feature is the key obstacle to the use of neural networks for medical decision making. If prediction is the only objective, then ANN models provide acceptable results, whereas logistic regression could also recognize the effect of factors on the classification (Kazemnejad *et al.*, 2010). Another hindrance of ANN model is the non availability of variable selection procedures unlike the logistic regression (forward elimination, backward elimination). ANN models have no comparable variable selection methods. But few researchers have suggested some ideas to explore effective factors in the model. West *et al*.

(West *et al.*, 1997) suggested that ANN model can be combined with other statistical models to reduce the number of input variables. Another way to reduce number of input variables is by examining the connection weight from the ANN model. Similarly, variables with low connection weights may be eliminated. Thus model should execute several times by inclusion or exclusion of variables and performance should be assessed (Larasati *et al.*). A very recent study suggested that the ratio of sum of square of errors with and without a given input variable should be assessed and if the value found is ≤1 then that variable may be discarded. This is a tedious and time consuming procedure and do not possess and ultimate solution. On the other hand, model is performed with the variables already shortlisted by univariate logistic regression and checked the overall performance; including sum of square of errors, correct classification, area under the ROC curve etc. With the same mode, settings and specifications ANN model is replicate several times to select only that one elucidating overall better performance.

Detailed analysis of advantages and disadvantages of each model is avoided, however, a summary comparison of each model is entailed in **Table 3.2**. No doubt, LR is a widely accepted method for modelling of binary response which should not be disregarded (Wong *et al.*, 2003), however, it has still certain limitations. All major limitations can be compensated by ANN model except; (1) proper interpretation of synaptic weights, (2) The type of association between predictors and outcome is not explicit in ANN model, (3) There is no proper variable selection method, (4) The order of variable inclusion might change the output, and (5) Calculation of odds ratio is not straight forward but is possible by means of calculating an "effective odds ratio" (Green *et al.*, 2006).

The two models may complement the other, emphasising different features of the data. Another important point with LR is that sometimes, it may not capture all possible significant covariates (Bourquin *et al.*, 1998) with the use of available variable selection methods and if anyone is conscious to explore all possible risk factors, the combine use of LR and ANN models can be a useful source. It will give the most accurate and comprehensive selection indeed. For example, in recent study, 11 risk factors are found to be significantly associated with the HCV outcome variable while one important variable "Number of persons per room" is additionally captured by ANN model which is actually fail to spot by LR, adding and making overall 12 risk factors. It is, therefore worth mentioning that if data analyzed for risk factors identification with LR model only then it is likely that some important risk factors might be ignored. This signifies the use of ANN model and can detect complex hidden patterns very easily within the data which is not apparent using traditional

statistics (Tabaton et al., 2010, Liew et al., 2007). This ensures that exact knowledge about all possible significant risk factors is very important for making preventive strategies. Moreover, it can be said that a great care and time are required to model the LR with all formalities and procedures. While on the other hand, ANN model can be a quick and easy multivariate analysis alternate to LR (Linder *et al.*, 2006).

In summary, the results showed that both-models performed almost equally, however, sensitivity and specificity of ANN model is slightly increased in training and testing sets of data. This increase may be due to inclusion of interaction effects in the model.

5.4 Modelling and Analysis of Risk factors Using Classification Tree Method

As stated earlier in section 2.5 that Classification Tree (CT) methods are useful alternative of widely used logistic regression. These are non-parametric, data mining, statistical methods and preferred due to ease of usability without getting any distributional assumptions. The popularity of such models is due to their visual demonstration of data making more understandable interpretation of the results. Another important aspect of Classification Tree method seemed to be powerful tool for identifying interactions between different risk factors which usually remains impractical in Logistic Regression due to complexity of the approach. It is known that risk factors interact with each other to ascertain disease risk (Camp and Slattery, 2002a) but the available literature regarding HCV infection looks silent in assessing interactions of risk factors. Therefore, complexity of interrelationships of risk factors for HCV infection is required attention and clarity. An application of Classification Tree method can be introduced to assess the pertinent risk factors of HCV together with multilevel interactions. Comparable to LR method, CT models can provide multivariate analysis and possess the ability to model categorical and also continuous dependent variable influenced by various predictors.

This study also aims to evaluate risk factors of HCV in Punjab, Pakistan using Classification Tree method and contrasting results with traditional Logistic Regression method. Moreover, this kind of analysis would certainly encourage the researchers and statisticians who are supposed to have little or no experience with the CT approach.

5.4.1 Development of CART Model for Overall Data

Although CART model has many limitations but still its importance can never be ignored. It's a relatively new approach for modeling binary response and researchers have applied in different fields even for exploring risk factors of a disease (1998). The CART

model uses a form of binary recursive partitioning and handles both continuous as well as categorical dependent (Target) variables. Importantly, the independent variables (predictors) of any measurement type can be entertained through this model. If the Target variable is taken as continuous, the method is referred as "Regression Tree" and if this target variable is measured as categorical variable it is known as "Classification Tree". In this study, our dependent variable is binary with "1" representing a Hepatitis C case and "0" a control.

The analysis is performed using IBM SPSS Decision Trees module. Different CART models are performed assuming different approaches to cater best model for identification of risk factors. The model is run with pre-defined specifications: maximum tree depth 5; the minimum numbers of cases (observations) in Parent node and Child node are taken as 100 and 50, respectively (Rastegari *et al.*, 2013). Several criterion functions are now available as impurity measures but Gini-Index is most widely used. This helps to identify best predictor and the best split to minimize impurity or in-homogeneity of the Child node (Nagy *et al.*, 2010). Minimum change in improvement is considered as 0.0001 and its large value tends to yield smaller tree.

To achieve an optimum tree model, different CART models are run on the same data set opting different criterions but presented only one having overall good performance **Figure 5.3**. Initially, the CART model is run considering all variables under study as independent variables and using aforementioned criterion. Later on, just like Logistic Regression model, CART model is run with only those independent variables having p-value<0.20 in univariate analysis as explicated in **Table 4.1** & **Table 4.2**. This is done because a CART model may include some un-important variables in the tree model or may exclude some important variables. Thus variables having potential influence in univariate analysis may restrict the model to have in-significant variables in the model.

The entire data set (n=1400) is utilized in growing the final tree model. However, validation is also attempted using split-sample validation method but results are not improved because sample size become reduced into two randomly generated samples i.e. training samples (80%) and tested on a hold-out sample (20%). This validation allows assessing how well tree structure generalizes to a population but could not be accompanied so far.

5.4.2 Analysis and Interpretation of CART Model

Figure 5.3 signifies a Classification Tree model for the final output. This tree model contains 15 nodes and 5 levels. These nodes comprised of 7 child nodes and 8 terminal nodes. The nodes are represented by the boxes and top node referred as Root node. This particular node includes all observations and implies the outcome variable. The outcome variable has two categories i.e. controls and cases with both percentage and numbers of how the sample divides between these categories. An annotated bar chart also describes percentage comparison of each category in each node. Other types of nodes including Parent nodes, Child nodes and Terminal nodes are also confessed on the diagram. When a node subsequently divides to create a new level it is called a Parent node. The analysis begins through the Root node and then splits by an independent variable. All of the independent variables are tested for their significance in splitting the data and only the most important predictor is captured at the 1st level. In this analysis, patients' education is determined as providing top ranked variable which splits into two categories, producing two child nodes. Initially patients' education is treated as a continuous variable but it splits into two categories i.e. <=8.5 years and >8.5 years by the CART model. CART model has ability to automatically categorize a continuous variable into binary splits very intelligently thereby no need of categorizing the continuous variable by the researcher. Of course, this is an amazing feature of CART technique. Subsequent data showed that about 62.7% cases have education <=8.5 years while (84.2%) majority of the controls are educated. Thus patients owning low education tend to be at higher risk. The comparison is rather distinct in the bar chart as well. Similarly, out of 33 predictors as recognized through the univariate analysis; only 6 are the most influential factors of hepatitis C risk within 5 levels of tree. Patients' education is found significant at two levels (level 1 and level 3). In CART analysis, a similar variable might be chosen repeatedly being an influential factor upon different levels of the tree. Other substantial factors with their level of importance are arrived at: local migration (Level 2), family history of hepatitis (Level 3), no. of persons sharing the room (Level 4), Major/Minor surgery (Level 4). Finally, history of blood transfusion is selected as a potential risk factor of hepatitis C at level 5.

Left child node of the first selected factor i.e. patient's education (<=8.5 years) is even further divided up by helping of a second discriminator since this node is not homogenous enough Three hundred and twenty one individuals of 1021 low educated (<=8.5) patients are allocated to the right child node, 700 patients having low education are allocated to the right

child node. Conversely, patients having low education, 81.6% cases have history of local migration/travelling. Thus the risk of HCV infection is higher in patients having history of local migration or travelling than others. The tree advanced one more step emanated from node 1 because of in-homogeneity and its child nodes divide into two independent variable. Left node inferred 3rd most important variable i.e. Family history of hepatitis, whereas patients' education cut-off value discriminating best in this step is 4.5 years. About 89.6% cases had education less than 4.5 years (Node 7). This again manifests that risk of disease further increases with decreasing years of education and adaptation of local migration/travelling. Family history of hepatitis is also an important risk factor of HCV infection at level 3. Out of 700 patients who had no history of local migration/travelling, 158 had positive history of hepatitis (Node 6). On percentage comparison, about 72.8% cases had positive history of hepatitis in this partition.

The factor "family history of hepatitis" does not possessing homogeneity so far and further partitioned into two more independent variables i.e. people sharing the room; major/minor surgery. These factors are emerged as level 4 factors both under the child nodes of family history of hepatitis. The No. of persons sharing the room's cut-off value discriminating best in this partition is 7.5, indicating the risk of disease increases in patients who had or presently sharing room with at-least 8 persons (Node 10). Similarly, the patients having positive family history and invasive surgery at once are at higher risk of disease. Last but not least, history of blood transfusion is also emerged as an important risk factor at 5th level of the tree model.

5.4.3 Performance of the Fitted CART model

Similar to the LR and ANN model, the performance of the CART model is evaluated by the ROC analysis. Area under the ROC curve is also computed along with its asymptotic 95% confidence interval i.e. AUROC=80.8% (95% CI: 0.785 to 0.830). Conversely, Classification table indicates that by using 6 variables about 74.5% of all patients can be correctly classified as to their cases/control status **Table 5.1.** The CART model thus constructed is the best fitted model having overall very good performance.

Table 5.1: Classification Table obtained after fitting CART model

Classification			
Observed	Predicted		
	Controls	Cases	Percent Correct
Controls	524	176	74.9%
Cases	182	518	74.0%
Overall Percentage	50.4%	49.6%	74.5%

Figure 5.2: ROC Analysis for CART Model

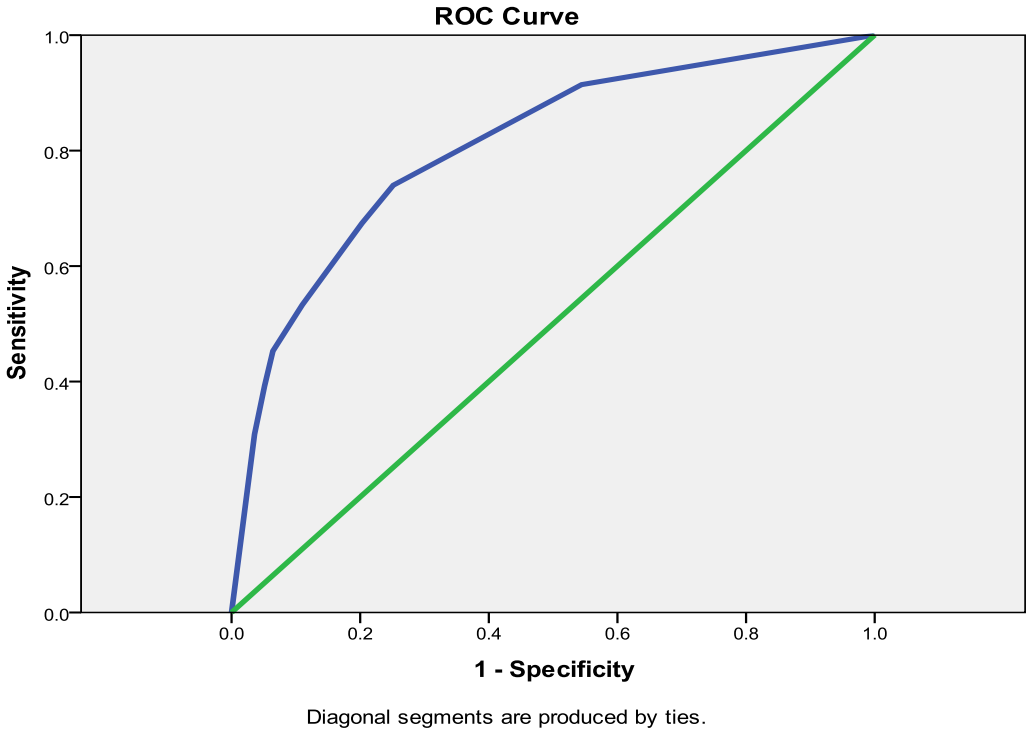
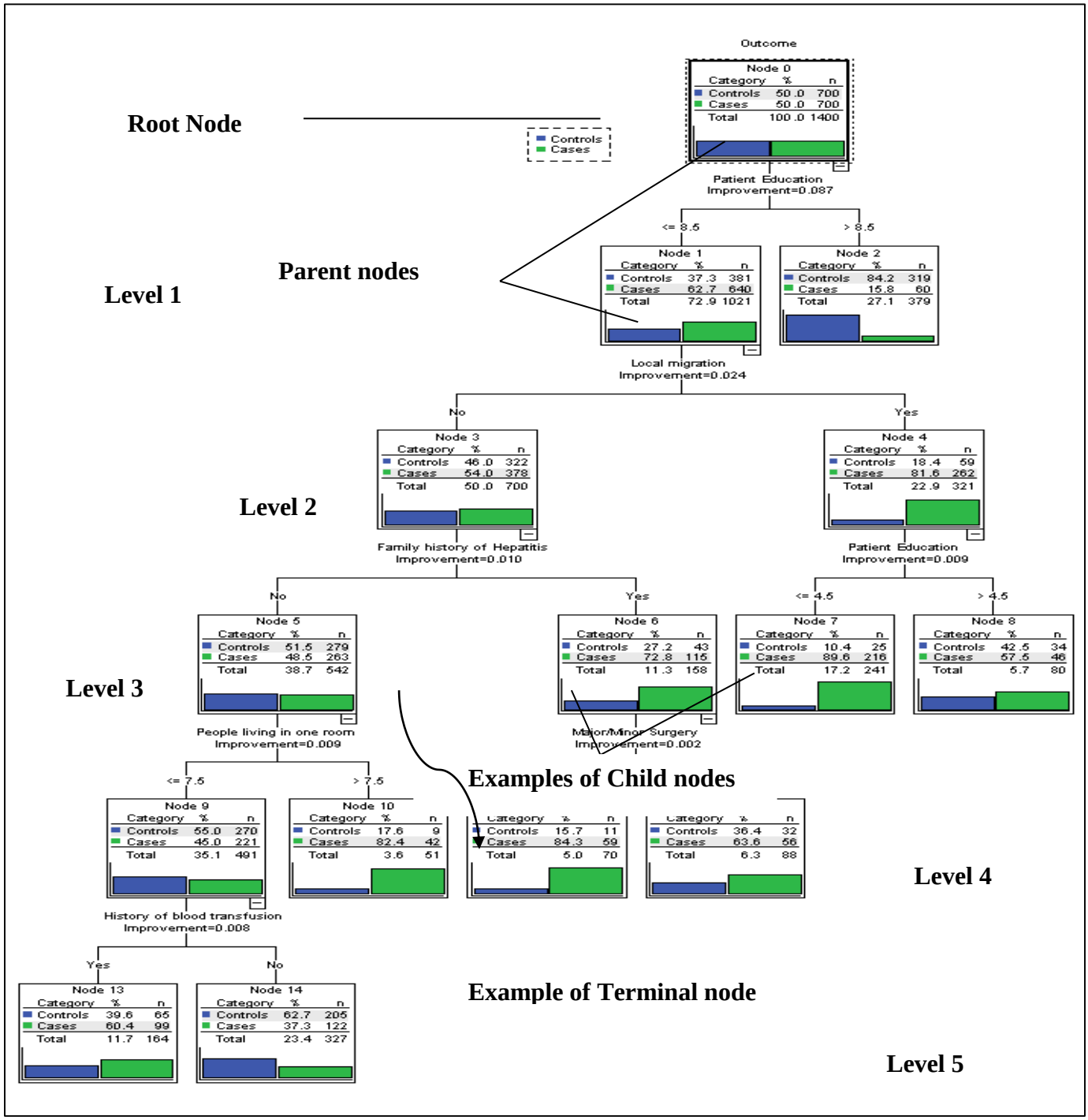


Figure 5.3: The CART Model Showing Potential Risk Factor in order of Importance



5.4.4 Comparison of CART and LR models

On comparison, 11 potential risk factors are identified in multiple logistic regression model from overall data, explicated in **Table 4.5** whereas only 6 risk factors examined in CART model effect hepatitis C disease significantly. It is found that five risk factors being found significant in Logistic Regression model are also found significant in CART model. One factor i.e. persons sharing the room could not be captured in LR model but selected as potential risk factor in the CART model. This reflects that as the number of persons sharing the room increases the risk of HCV disease increases and vice versa. Although CART model has captured smaller number of risk factors in the tree structure compared to Logistic Regression model but both models hold almost match-able overall correct classification to differentiate between cases and controls i.e. approximately 75%. In spite of this, misclassification of controls/cases is found bit larger in CART than LR model. Moreover, considering the whole data set, AUROC curve is 80.8%, 83.8% in CART and LR model respectively. Overall both models have indicated very good fit. Additional number of potential risk factors can also be added in the CART model by enhancing tree levels but tree structure turned complex with larger number of variables. As soon as tree grows up, sample size in several terminal nodes reduces. These limitations are acceptable to ascertain the hypothesis and the usefulness of the tree model. Nevertheless, pruning of the tree is also promising subject to contentment of the researcher.

The results obtained from CART model are more detailed and explain structure of risk factors in a better way. The results also showed interactive behavior of the risk factors in 5 different levels of importance which manifests that many factors at many levels are involved in epidemiology of HCV disease. Thus, CART model assists the assessment of interactive effects of variables. In the present study, for example, local migration/travelling is the important predictor of HCV infection among the subjects who have less than 8.5 years of education. Subjects having low education often had to migrate for their livelihood and travel for getting better earning opportunities. For instance, patients in driving occupation often have to travel frequently and stay at distant locations for short span of time with miserable living facilities in fact. But, among the patients who have not history of frequent travelling/ migration showed interaction with “Family History of Hepatitis”. This interacting behavior clearly explains that the subjects who have started to live at the same locality and do not

migrate/travel in frequent, involve in the habit of sharing of daily items which ultimately augment the risk of HCV disease. For example, habit of sharing of; nail clipper, needles, razors and tooth brushes, appear with time. Other multilevel interactions are also apparent from the tree structure which is impractical in Logistic Regression model due to complexity of approach (Camp and Slattery, 2002a).

Possibly, classification tree methods work efficiently to identify the correct combination of interactions and to understand disease pathways but this is not true all the time. Therefore, it is suggested that different trees be investigated on the same data set to reach on most practical and clinically established interaction effects. In this study, multiple classification trees methods are not attempted as the objective of this study is to explore the usefulness of the approach. Certainly, results are encouraging and informative in modeling and analyzing the risk factors of HCV infection using such a powerful data-mining technique. This really helps to explore complexities of disease risk in a simple and promising way.

Table 5.2: Summary of LR, ANN and CART Models on Overall Data

Model / **Parameter**	**Logistic Regression**	**Artificial Neural Networks (ANN)**	**CART**
Potential Variables selected	i. Patient's education ii. Local migration/Travelling iii. H/O of liver disease iv. Endoscopy v. Family H/O hepatitis C vi. Tattooing vii. Blood Transfusion viii. Minor Surgery by Barber ix. Dental Surgery x. Major/Minor Surgery xi. H/O injections	i. Patients education ii. H/O liver disease iii. Local migration/Travelling iv. Family H/O hepatitis C v. No. of persons sharing the room vi. Blood transfusion vii. Endoscopy viii. H/O injections ix. Dental Surgery x. Tattooing xi. Major/Minor Surgery	i. Patient's education ii. Local migration/Travelling iii. Family H/O hepatitis C iv. No. of persons sharing the room v. Major/Minor Surgery vi. Blood transfusion
Correct classification	75.4%	Training data (77.3%) Testing data (79.8%)	74.5%
AUROC	83.8%	84.7%	80.8%
Sensitivity	76.1%	Training data (77.6%) Testing data (77.0%)	74%
Specificity	74.8%	Training data (78.5%) Testing data (81.2%)	74.86%

Table 5.3: Showing Summary/Comparison of Significant Risk Factors Identified in Different Models

Models on overall data	**Using Logistic Regression Models**			
Logistic Regression (Gold Standard)	**Overall-Male**	**Overall-Female**	**North Punjab Model**	**South Punjab Model**
Patient's education	No. of persons per room	No. of persons per room	Gender (Male)	Patient's education
Local migration/Travelling	Patient's education	Patient's education	No. of persons per room	Patient history of liver disease
H/O of liver disease	Family H/O Hepatitis C	Family H/O Hepatitis C	Patient's education	Blood Transfusion
Endoscopy	Blood Transfusion	Patient history of Jaundice	Family H/O Hepatitis C	H/O blood donation
Family H/O Hepatitis C	H/O blood donation	Blood Transfusion	Patient history of Liver disease	Tattooing
Tattooing	Dental Surgery	Ear/Nose piercing	Blood Transfusion	Sharing of tooth brush/Miswak
Blood Transfusion	Nails cut from barber	Injected drugs, even once	Dental surgery	History of endoscopy or gastroscope
Minor Surgery by Barber	Local migration/travelling	Major/Minor Surgery	Injected drugs, even once	
Dental Surgery	Minor surgery by barber	History of abortion/D&C	H/O accidental needle stick	
Major/Minor Surgery		Road crash injury	Major/Minor Surgery	
H/O Injections		Hospitalization	History of injections/Intravenous drips	
		Local migration/Travelling	History of branula insertion	
			Minor surgery by barber	
ANN Model	**Urban-Male**	**Urban-Female**	**Male**	**Male**
Patient's education	No. of persons per room	No. of persons per room	No. of persons per room	Patient's education
H/O liver disease	Patient's education	Patient's education	Patient's education	Patient history of liver disease
Local migration/Travelling	Blood donation	Family H/O Hepatitis C	Family H/O Hepatitis C	Blood Transfusion
Family H/O Hepatitis C	Dental surgery	Ear/Nose piercing	Blood Transfusion	Nails cutting from barber
No. of persons per room	Tattooing	Injected drugs, even once	H/O blood donation	Removal of un-wanted hairs from the "Hammam"/hostels or common flats
Blood transfusion	History of needle click	Accidental needle click	Dental Surgery	History of endoscope or gastroscope
Endoscopy	Sharing of razors for shaving	History of abortion/D&C	Sharing of Razors	Minor surgery by barber
H/O injections	Nails cutting from barber	History of branula insertion	Major/Minor Surgery	

Dental Surgery	Removal of un-wanted hairs from the "Hammam"		History of injections/Intravenous drips	
Tattooing			Minor surgery by barber	
Major/Minor Surgery				
CART model	**Rural-Male**	**Rural-Female**	**Female**	**Female**
Patient's education	Patient's education	No. of persons per room	No. of persons per room	Family status (Nuclear family)
Local migration/Travelling	Family H/O Hepatitis C	Patient's education	Patient's education	Patient's education
Family H/O Hepatitis C	Blood transfusion	Blood Transfusion	Family H/O Hepatitis C	Family H/O Hepatitis C
No. of persons per room	H/O blood donation	History of abortion/D&C	Patient H/O Jaundice	History of accidental needle
Major/Minor Surgery	Nails cut from barber	History of endoscope or gastroscope	Blood Transfusion	History of abortion/D&C
Blood transfusion	Minor surgery by barber	Have you had passed hostel life?	Ear/Nose piercing	
	Have you had passed hostel life?		Sharing of nail cutter	
			Major/Minor Surgery	
			History of abortion/D&C	

Chapter 6

Summary, Conclusions & Recommendations

In this Chapter a brief summary of the whole thesis is given which comprises of main objectives and how these are achieved by applying aforementioned methodology. A summary of main findings along with important conclusions is also described. Additionally, some limitations of the study, recommendations for prevention of HCV and recommendations for future research are also included in separate sections.

6.1 Summary

It is a famous saying that **"Prevention is better than cure"** however, in order to prevent a particular disease, one must know the exact routs, causes or ways of transmission of that disease. Hepatitis C is a chronic disease which is caused by the HCV virus. Prevalence of HCV infection in Pakistan is estimated at 4.9%, while in Punjab it is 6.7%. Global prevalence of HCV infection is also increasing rapidly. No vaccine is being offered currently for this disease. Therefore, many experts focused to study epidemiology of HCV with an aim to identify its pertinent associated "Risk factors". Literature regarding HCV prevalence and its epidemiology has demonstrated that both are country-specific. A fair research work regarding risk factors of HCV infection is available in developed countries, but it is not true for developing countries like Pakistan. Also modes of transmission vary by region. Further its treatment requires high expenses. Pakistan being the developing country is struggling to meet this challenge with meager resources though with little success. Thus, it is essential to identify how the virus is transmitted in order to plan preventive measures. Logistic regression has traditionally been used to determine risk factors but there are now alternatives available, such as Artificial Neural Network (ANN) and CT Models, which have gained attraction in recent years. This study provides a compact form of research work wherein application of parametric (logistic regression) as well as non-parametric (Neural networks and Classification Trees) statistical techniques are applied to an epidemiological study. Conversely, this study would present an example of how the **LR** (Gold standard), **ANN** and **CT** modelling techniques can be used to study risk factors for hepatitis C infection. The aim

of study is not to establish superiority of any of the method but to explore, compare and contrast the results of each method.

A consecutive sample of 1,400 patients is collected from the patients of nine different Divisional Head Quarter (DHQs) hospitals of the Punjab Province, Pakistan with 700 cases and 700 controls. The cases are HCV positive, diagnosed by the ELISA method; whilst the controls are negative. The patients are recruited from outpatient, inpatient and hepatitis clinic (where applicable) settings from each hospital. Risk factors and socio-demographic information are sought using an administered questionnaire. Associations between risk factors and disease status are compared and contrasted using LR, ANN and CT Models. All these models have the ability to predict binary outcomes (Yes/No) with many predictors.

The descriptive statistics from the case-control study showed that out of a total of 1400 patients, 761 (54.4%) are male and 639 are (45.6%) female. The mean (±SD) age is 34.1±10.5. The range of the patient's ages lies between 14-80 years. Overall 78.9% are married, 60.6% of subjects reported living in a rural area, whilst 52.4% lived with an extended family in one dwelling. About 68.7% cases reported having no education, whilst others are generally low. Monthly household income is similar for both (88%) cases and controls, with the majority belonging to a low income group (<15000 Rs./month or 160 US$/month). On average, participants lived in a family of 7 to 8 people; 3 people shared one room (in controls) and 5 persons in cases. Among the cases, about 68.7% patients' fathers had no education, and this figure increased in mothers (78.0%).

The significance of this study is a comprehensive examination of risk factors that has been performed by using different Logistic Regression models in various settings of data, for example, overall data, male/female, urban/rural and geographical settings to identify pertinent risk factors at their exact place. Before development of the multivariate logistic regression, univariate analysis is executed at first and only the predictors owning p-value<0.2 are chosen for final selection in the multivariate Logistic Regression Model. Diagnosis of multicollinearity and outliers is also done just before finalizing the model. But analysis against other techniques such as ANN and CT models is solely performed for overall data, avoiding extended explanation. This investigation offers the basis for comparing results in ANN and CT models with LR (Gold standard) model. Analysis is performed using IBM SPSS version 19 at University of Auckland, New Zealand.

On comparison, results from LR and ANN models which are built on overall data, matching risk factors are identified by both models, apart from one variable **Table 5.2**. On combining the results from both models, 12 risk factors are strongly associated with HCV infection. It can be concluded that a concordant set of risk factors, from both LR and ANN models is identified which implies that ANN is a useful adjunctive method to identify risk factors for Hepatitis C infection in Punjab, Pakistan. Overall both models equally performed but figures in correct classification, AUROC, Sensitivity and Specificity are relatively improved in ANN model than LR. This may be due to the inclusion of interaction effects in the ANN model. Despite the ANN model including interaction effects in the model, it is challenging to convey the meaning of these model characteristics. Only networks can recognize these interactions and their effects are evaluated in the model intrinsically. If someone is curious to watch these interaction effects, Classification Tree models are helpful which allow the discovery of pertinent factors with their multilevel interactions. These are generally easy to understand in the tree structure of the CT model . In this study, CART model has also been applied and experienced very explanatory role in identifying essential risk factors and even their interactions. Compared to LR, the CART model showed the similar figure of correct classification but with slightly lesser AUROC. Moreover, this CART model showed that only 6 risk factors i.e. Patient's education, Local migration/travelling, Family H/O Hepatitis C, No. of persons sharing the room, Major/Minor surgery, and receipt of a blood transfusion are strongly associated with HCV infection. These six risk factors have also been identified in LR as well ANN models. Interestingly, when all significant risk factors obtained through different settings/models of patients are summarized/compared in **Table 5.3** the same 6 risk factors are turned out as most consistently identified risk factors. Hence, these can be declared as the most common and general risk factors of HCV infection in the Province. Almost every model revealed that subjects who received formal education are less likely to be HCV positive. This indicates that as the level of education increases, the understanding and awareness among the people affected in a positive way (Asghar *et al.*, 2009). Other important risk factors in male patients with urban/rural settings of data are having had dental surgery, tattooing, sharing razors, habit of nails cutting from barber, minor surgery by barber, removal of unwanted hairs with communal razors placed in "Hammam"/hostels/flats common wash rooms and having lived in a hostel at some point in the past. Odds ratios which describe their estimated level of risk are also computed in each Logistic regression model along with their 95% CIs. This can be the important aspect of LR analysis of which relative risk of each significant risk factor can be characterized by

computing odds ratio which may distinguishes it from the ANN and CART models. Likewise, among the female patients with urban/rural settings of patients, some different risk factors are identified in this study: Patient H/O Jaundice, Ear/Nose Piercing, history of abortion/D&C, Road crash injury, Hospitalization, accidental needle click and Branula insertion. These are expectedly different risk factors compared to males because of their differences in biological, lifestyle/behavioral and medical history related factors.

Similarly, on regional classification of data, different LR models are run for overall, male and female data sets in North and Southern regions of Punjab independently. In North region of Punjab, being male, H/O liver disease, dental surgery, injected drugs, H/O accidental needle stick, cannula insertion, minor surgery by barber are the major risk factors in overall data. Correspondingly, in South region, some additional risk factors such as H/O liver disease, H/O blood donation, tattooing, sharing of tooth brush/miswak and endoscopy are standing out as major risk factors which are certainly different compared to North regions of the Province. Gender-wise comparison of risk factors in both regions is made which reflects that some risk factors within male/female models also differ considerably **Table 5.3**.

Thus, after making this detailed analysis, some risk factors differed by gender, residential location and region. This supports our hypothesis and concludes that researchers have to explore different set of risk factors by examining different strata of the data. This would help to understand the relevant risk factors at their best place. This way to partition the data would definitely help to introduce the targeted preventive campaigns in the community. It is worth mentioning that out of total 55 variables taken as suspected risk factors, 36 are emerged as potential source of infection in the Province in different settings/arrangements of data.

In summary, it is to be noticed that certain potential risk factors belong to socio-economic, medical history, behavioral characteristics and family history of the patients. But majority of the risk factors are related to patients' medical history which highlight that major factor for HCV infection in Punjab is due to a lack of hygiene and sterile technique, applied by a variety of hospitals/clinics, when receive healthcare. Given the concordance of recent study with other studies carried out in Pakistan, and in other developing countries, the risk factors have been identified are likely to be causal, and warrant the initiation of preventive action. It is suggested that an audit of infection control procedures be carried out in all medical settings in which blood-to-blood contact may occur, and that only those institutions that meet minimum requirements be allowed to continue to practice. Practical and cost-

effective means of maintaining sterile instruments are urgently required to prevent infection. Endoscopy, dentistry, transfusion services and surgical /gynecological procedures stand out as priorities.

6.2 Conclusions

The primary objectives of this study are to assess significant risk factors for HCV and develop statistical models for its occurrence with a view to using the findings to identify individuals at risk in Punjab. In addition, this study would also draw the attention of the researchers to look into other techniques besides conventional methods in modeling and analyzing risk factors of hepatitis C disease for better insight. Based on the objectives of the study, the first objective is to apply descriptive and analytical analysis for describing risk factor in better spirit. An in depth descriptive analysis coupled with proper discussion on each risk factor is accompanied in the study. Analytical section of the study is regarded as the important segment of the study consisting of univariate and multivariate analysis. This analysis really helps to assess the association of each risk factor with the disease outcome. The Chi-square test and Fisher's Exact Tests are performed with univariate logistic regression model as univariate assessment statistical methods. Afterwards, multivariate analysis with logistic regression model is carried out on overall data to undertake 2nd most important objective of the study. This analysis identifies that total 11 risk factors are potential or significant in overall model, also given odds ratios and their 95% CIs. Odds ratios would be the useful tool for interpreting the risk factor in terms of likelihood or risk of the disease.

To be able to assemble 3rd objective of the study, separate logistic regression models are built on male/female data. It was done due to the differences in biological, behavioral and exposure to medical treatment history related factors in both genders. The discussion hence integrated in summary section 6.1 implies that certain factors vary in male and female settings of data. Similarly, to fulfill the 4th objective, separate LR models are constructed for urban/rural settings of patients which highlighted that a number of factors differ during these settings when models were performed independently.

To include regional classification of data, region-specific models are developed for the patients who belong to Southern and Northern areas of the province Punjab, independently. This is actually the 5th objective of the study. These partitioning of data facilitated us to learn even more hidden risk factors which could not be discovered in the prior models. Since the lifestyle, behavioral characteristics and health facilities are different

in both regions significantly, therefore, some risk factors are again distinctive in these regions.

Last but not the least, 6th major objective of the study is introduced by applying ANN and CART model on overall data to explore pertinent risk factors. The application of these non-parametric, data mining techniques is rare but applicable in modeling and analyzing the risk factors data. Thus, all enlisted objectives are achieved satisfactorily and the results reveal that a variety of exposures from the socio-demographic factors, behavioral characteristics of patients, medical & health care procedures and community exposures are associated with hepatitis C infection in Punjab. Also, it is noticeable that risk factor behaviors vary by gender, residential location and region of the patients which satisfies our hypothesis statement. The results also suggested that by applying LR, ANN and CT models, the researchers are better able to explore those risk factors which can increase susceptibility to disease. These findings are likely to improve preventive strategies and awareness of how HCV infection spread in the community. In short, this also implies that researchers should not rely only on traditional methods to model and analyze the risk factors of HCV infections but also make use of other techniques such as artificial neural network and CT together for better insight to the data exploration, more refined understanding on the main determinants of HCV infection.

6.3 Recommendations for Prevention of HCV

Prevention is only possible when exact routes or causes of disease are well-known. An in-depth analysis of risk factors as rendered in this study may help to direct effective policies and then to disseminate awareness of routes of transmission among the general public. This tends to diminish the burden of morbidity and mortality due to hepatitis C. Some useful measures are listed below which can help to eradicate this disease from the region.

- Education for all should be the top priority of the Government. It will prove to be helpful to prevent and combat this disease.
- Stringent legislation for the purpose of licensing and inspection of medical and dental practitioners, barber shops, prostitutes, tattooing, acupuncture and ear-piercing should be made and implement in larger public interest.
- Although free treatment of hepatitis C is available in DHQs Hospitals at certain extent but it should be readily available in all government hospitals.

- Medical practice by unqualified practitioners should not be legislated against and only qualified/trained health care workers be recruited immediately, particularly in rural areas.
- Implementation of proper sterilization and disinfectant methods in Government hospitals should be enacted.
- It was observed that waste material that contain used syringes, surgical appliances and laboratory kits are being dispersed around the hospitals and not properly disposed-off/dumped. Therefore, a safe waste disposal system should be immediately implemented in government hospitals.
- One of the major causes of HCV infection is the use of unsafe blood transfusion. To cap this situation, the same should be avoided and only the screened blood be transfused, if required. Paid blood donors should be discouraged. Poor or substandard screening kits are being utilized in Pakistan which requires proper legislation to ensure blood is properly screened for infection before transfusion.
- Students in schools/colleges/universities may be a useful source to spread awareness among the population, particularly to family members who often remain unreached. But, firstly, we need to educate them about the precise knowledge of the risk factors and then they would be better able to contribute.
- The study implies that majority of the factors originates due to the medical treatment history in public sector hospitals. Hence, an adaptation of **sense of responsibility** in doctors, nurses and other paramedical staff is obligatory who are supposed to know well about the risk factors of this disease. They should ascertain proper sterilization of the instruments before performing endoscopy, gastro scope or angiography, surgical operations, dental surgery, cesarean, gynecological procedures and blood transfusion. A minor negligence at their part can lead to this dangerous disease. Simply, the active and responsible role of medical personals' may minimize the risk.
- Now a day's access to print and electronic has become possible even in remote areas of Pakistan. However, unfortunately, the role of the media in creation of awareness is not up to the required level. If used effectively, it may be the cheapest method to prevent viral transmission.
- The results reveal that majorities (61%) of the patients belong to rural areas where health-related facilities are limited compared to urban areas. Eventually, patients often had to travel long for their better treatment, otherwise, they simply rely on locally available un-registered quacks for dental extraction, minor surgery and other medical

treatment. Minor sugary or cut on wound is additionally common by the barber which explains its association with the disease. Females also had to visit those non-professionals, untrained midwifes or nurses for their deliveries or other gynecological procedures where there is no concept of sterilization. It is, therefore, recommended that proper health facilities be provided in the rural areas urgently with qualified and trained health care workers.

- Although private hospitals are compensating the burden of patients. But, here, qualified doctors charge high fees which poor people may not be able to pay. There should be legislation that doctors while doing their private practices could not charge such heavy amounts from the patients in the best interest of humanity.
- Public awareness campaign should be launched through different ways such as organizing workshops, seminars, in local communities for the local people.
- Risk factors which are related to behavioral characteristics of the patients i.e. tattooing, ear-nose piercing, sharing of tooth brushes, nail cutter sharing, barber shaving/armpit, communal razor for removing unwanted hairs etc can be stopped by simply educating the people.
- Persons having extra-marital relationship should follow safer ways including contraceptive.
- The researcher observed that about 90% of People with known HCV infection still have no awareness about the causes of this disease. It is, therefore, recommended that these patients should be counseled properly to reduce the risk of transmitting to others. In this regard, guidance and counseling centers may be established in the hospitals where patients visiting for treatment may get proper information/ awareness about different diseases at any cost including hepatitis C.
- Government of Pakistan is spending millions of Rupees just for hepatitis C treatment. Pakistan is being the developing country cannot bear all of those expenditures. Hence, the prevention/controlling of this hazardous disease may conserve the resources and enhance working capabilities or the manpower. This ultimately leads in the direction of economic development of the country.
- In this study not a single patient have reported about health insurance. While, health insurance may prove to be helpful for resource-constraints patients. In this regard, it is recommended that affordable insurance policies be introduced with Government intervention enabling the patients to get rid this chronic disease without much financial burden.

- ❖ The Government of Pakistan should devise an effective policy for rehabilitation of Afghan refugees and IDPs (Internally displaced persons) of Northern areas and provide healthcare services to them on immediate basis including blood screening for hepatitis C infection, in particular.
- ❖ Last but not least, it should come realize that prevention is only the cost effective way to reduce the burden of disease. However, the responsibility does not lie with the Government solely. In order for the dream of healthier Pakistan to come true, it is imperative that civil society should come forward and play their role jointly in creating awareness among the masses about causes, remedies, and preventive measures of the disease.

6.4 Limitations of the Study

Although an attempt is made to ensure maximum possible accuracy in the study in terms of data collection, methodology and results, yet there are certain limitations given below.

- ➢ This study used hospital controls and relied on self-report data for the majority of risk factors reported (Khan et al., 2008b). The study is limited by resources, though it is tried to meet sufficient sample size as estimated. The use of hospital controls might have lead to distortion of effect estimates, given that many controls are probably visiting the hospital as a result of the symptoms of a range of other infectious diseases, which may share similar risk factors to those of HCV infection. It might be expected that this selection bias is likely to result in an under-estimate of measures of association. The use of prevalent, rather than incident cases, might further have distorted measures of association, due to the effect of survivor bias. The case-control design also may not distinguish cause from effect, so, for example, a diagnosis of HCV infection may increase the probability of medical procedures, such as gastroscopy or minor surgery.
- ➢ Overall sample size is adequately large; however decreased once subdivision or partitions are created to build up separate models for male/female patients. Later on, additional subdivision of data is continued under urban-male, rural-male; rural-male, rural-female settings of patients. Regional classification of data is also ascertained within this study but this also reduces the sample size. These partitioning of data had been essential to ascertain at present because no studies in current literature originate with such separate versions. Therefore, different models with different settings of information are performed, even though with small sample size, to search for pertinent risk factors at their best place.

- Un-safe marital relation is a suspected and reported risk factor of HCV infection in international literature but could not be studied in Pakistan due to societal constraints. This study also accounted for patients victimized this way, however, the subject area is kept confined to male patients only. On account of societal and religious constraints, female subjects could not be interviewed on these lines.

6.5 Recommendations for future Research

Some important recommendations for future studies have been listed below which can be useful for new researchers as well as policy makers.

1. Application of Artificial Neural Networks and Classification Tree methods may also be experienced on other data sets in order to determining risk factors of other diseases.
2. ANN and CT models can be performed with or without applying Logistic Regression in identification of risk factors but the use of these data mining techniques can be more fruitful, if applied with LR model.
3. Data should also be collected using better sampling design i.e. systematic, cluster or stratified random samplings. The use of **multilevel *logistic regression*** modeling technique can be introduced in risk factors identification by integrating the effects of different factors at different levels. These levels may be the patients' level, family level and community level.
4. A separate study may be conducted on children to determine risk factors of hepatitis C infection.
5. Risk factors of hepatitis C must also be investigated among the high risk groups i.e. blood donors, dialysis and drug using patients.
6. Risk factors of hepatitis C in diabetic patients may also be explored separately.
7. Measuring of Quality of Life (QOL) of hepatitis C patients should also be the focus of the researchers.
8. This study targeted the Punjab province only but it also necessitates conducting similar studies from the other Provinces of Pakistan.
9. Occupational risks/hazards of hepatitis C should also be studied in detail in separate studies.
10. Many of such case-control studies have been carried out using hospital based data but studies with community based data are lacking and asking for further attention of the researchers.

11. Prevalence of hepatitis C in Pakistan is not much clear, therefore, studies having good sampling design should also be ascertained in different regions of Pakistan.
12. Risk of hepatitis C due to history of unsafe/extramarital contact may also be investigated in depth in both genders/spouses, particularly in females as this area could not be investigated in females due to social constraints. Although prostitution is strictly prohibited both religiously and legally in Pakistan yet it is need to be taken into account for this subject area.
13. Although data in this study is collected in each division of Punjab Province yet it could not be analyzed separately avoiding details. Therefore, it is also suggested to apply such analysis with larger sample size in each division of Province Punjab.
14. A central data base in any field is of utmost importance as it does not reduce financial cost only but also helps the researchers/policy makers reach their goals in minimum possible time. Unfortunately, there is no such data base available in Pakistan regarding hepatitis C due to which the researchers have to invest huge financial and time costs for their data collection. In view of this, it becomes imperative that a central data base of hepatitis C patients must be available at national level and accessible to the indenting researchers, if and when required.

References

Abbas, Z., Jeswani, N., Kakepoto, G., Islam, M., Mehdi, K. & Jafri, W. 2008. Prevalence and mode of spread of hepatitis B and C in rural Sindh, Pakistan. *Trop Gastroenterol,* 29**,** 210-6.

Afonso, A. M., Ebell, M. H., Gonzales, R., Stein, J., Genton, B. & Senn, N. 2012. The use of classification and regression trees to predict the likelihood of seasonal influenza. *Family Practice,* 29**,** 671-677.

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*, Wiley.

Ahmad, N., Asgher, M., Shafique, M. & Qureshi, J. A. 2007. An evidence of high prevalence of Hepatitis C virus in Faisalabad, Pakistan. *Saudi medical journal,* 28**,** 390.

Ahmed, F., Irving, W. L., Anwar, M., Myles, P. & Neal, K. R. 2012. Prevalence and risk factors for hepatitis C virus infection in Kech District, Balochistan, Pakistan: most infections remain unexplained. A cross-sectional study. *Epidemiol Infect,* 140**,** 716-23.

Ahmed, I., Ali, Z., Alam, I., Nazir, S. M., Taqweem, A. & Mahboob, A. 2010a. Risk Factors Associated with Hepatitis C virus Acquisition in a Tertiary Care Setting. *Gomal Journal of Medical Sciences,* 9**,** 32-38.

Ahmed, I., Ali, Z., Alam, I., Nazir, S. M., Taqweem, A. & Mahboob, A. 2010b. Risk Factors Associated with Hepatitis C virus Acquisition in a Tertiary Care Setting. *Gomal Journal of Medical Sciences,* 9.

Akahane, Y., Kojima, M., Sugai, Y., Sakamoto, M., Miyazaki, Y., Tanaka, T., Tsuda, F., Mishiro, S., Okamoto, H. & Miyakawa, Y. 1994. Hepatitis C virus infection in spouses of patients with type C chronic liver disease. *Annals of Internal Medicine,* 120**,** 748-748.

Akbar, H., Idrees, M., Manzoor, S., Rehman, I. U., Butt, S., Yousaf, M., Rafique, S., Awan, Z., Khubaib, B. & Akram, M. 2009. Hepatitis C virus infection: A review of the current and future aspects and concerns in Pakistan. *J Gen Mol Virol,* 1**,** 12-18.

Akhtar, S., Moatter, T., Azam, S. I., Rahbar, M. H. & Adil, S. 2002. Prevalence and risk factors for intrafamilial transmission of hepatitis C virus in Karachi, Pakistan. *J Viral Hepat,* 9**,** 309-14.

Akhtar, S., Younus, M., Adil, S., Jafri, S. & Hassan, F. 2004. Hepatitis C virus infection in asymptomatic male volunteer blood donors in Karachi, Pakistan. *Journal of viral hepatitis,* 11**,** 527-535.

Alavian, S.-M. 2007. Control of hepatitis C in Iran: vision and missions. *Hepat Mon,* 7**,** 57-8.

Alavian, S. M. 2010. Hepatitis C virus infection: Epidemiology, risk factors and prevention strategies in public health in IR IRAN. *Gastroenterology and Hepatology from bed to bench,* 3**,** 5-14.

Alavian, S. M. & Aalaei-Andabili, S. H. 2011. More risk factors of hepatitis C transmission should be considered in Pakistan. *Int J Prev Med,* 2**,** 188-9.

Alavian, S. M., Einollahi, B., Hajarizadeh, B., Bakhtiari, S., Nafar, M. & Ahrabi, S. 2003. Prevalence of hepatitis C virus infection and related risk factors among Iranian haemodialysis patients. *Nephrology (Carlton),* 8**,** 256-60.

Alavian, S. M., Gholami, B. & Masarrat, S. 2002. Hepatitis C risk factors in Iranian volunteer blood donors: a case-control study. *J Gastroenterol Hepatol,* 17**,** 1092-7.

Ali, S. A., Donahue, R. M., Qureshi, H. & Vermund, S. H. 2009a. Hepatitis B and hepatitis C in Pakistan: prevalence and risk factors. *Int J Infect Dis,* 13**,** 9-19.

Ali, S. A., Donahue, R. M. J., Qureshi, H. & Vermund, S. H. 2009b. Hepatitis B and hepatitis C in Pakistan: prevalence and risk factors. *International Journal of Infectious Diseases,* 13**,** 9-19.

Allison, P. D. 1999. *Logistic regression using the SAS system: theory and application*, SAS Publishing.

Almeida, L. M., Werneck, G. L., Cairncross, S., Coeli, C. M., Costa, M. C. & Coletty, P. E. 2001. The epidemiology of hepatitis A in Rio de Janeiro: environmental and domestic risk factors. *Epidemiol Infect,* 127**,** 327-33.

Alter, H. J., Purcell, R. H., Shih, J. W., Melpolder, J. C., Houghton, M., Choo, Q. L. & Kuo, G. 1989. Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non-B hepatitis. *The New England journal of medicine*, 321**,** 1494.

Alter, M. J. Year. Epidemiology of hepatitis C in the West. *In:* Seminars in liver disease, 1995. 5-14.

Alter, M. J. 2003. Epidemiology of hepatitis C. *Hepatology,* 26**,** 62S-65S.

Alter, M. J. 2007. Epidemiology of hepatitis C virus infection. *World Journal of Gastroenterology*, 13**,** 2436.

Alter, M. J., Hadler, S. C., Judson, F. N., Mares, A., Alexander, W. J., Hu, P. Y., Miller, J. K., Moyer, L. A., Fields, H. A., Bradley, D. W. & Et Al. 1990. Risk factors for acute non-A, non-B hepatitis in the United States and association with hepatitis C virus infection. *JAMA*, 264**,** 2231-5.

Alter, M. J., Kruszon-Moran, D., Nainan, O. V., Mcquillan, G. M., Gao, F., Moyer, L. A., Kaslow, R. A. & Margolis, H. S. 1999. The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. *New England Journal of Medicine*, 341**,** 556-562.

Apolloni, B., Marinaro, M. & Wirn 2002. *Neural Nets: Proceedings of the ... Italian Workshop on Neural Nets. Vietri sul Mare, Italy, May 30 - Juni 1, 2002*, Springer.

App. 2012. *150 hospitals to have free hepatitis treatment facility* [Online]. Available: http://app.com.pk/en_/index.php?option=com_content&task=view&id=10088 [Accessed].

Armstrong, G. L., Wasley, A., Simard, E. P., Mcquillan, G. M., Kuhnert, W. L. & Alter, M. J. 2006. The prevalence of hepatitis C virus infection in the United States, 1999 through 2002. *Annals of Internal Medicine,* 144**,** 705.

Asghar, Z., Attique, N. & Urooj, A. 2009. Measuring Impact of Education and Socio-economic Factors on Health for Pakistan. *The Pakistan Development Review*, 48**,** pp. 653-674.

Ashraf, H., Alam, N. H., Rothermundt, C., Brooks, A., Bardhan, P., Hossain, L., Salam, M. A., Hassan, M. S., Beglinger, C. & Gyr, N. 2010. Prevalence and risk factors of hepatitis B and C virus infections in an impoverished urban community in Dhaka, Bangladesh. *BMC Infect Dis,* 10**,** 208.

Aslam, M. & Aslam, J. 2001. Seroprevalence of the antibody to hepatitis C in select groups in the Punjab region of Pakistan. *Journal of clinical gastroenterology*, 33**,** 407.

Averhoff, F. M., Glass, N. & Holtzman, D. 2012. Global burden of hepatitis C: considerations for healthcare providers in the United States. *Clinical infectious diseases*, 55**,** S10-S15.

Awadalla, H. I., Ragab, M. H., Nassar, N. A. & Osman, M. A. 2011. Risk factors of hepatitis C infection among Egyptian blood donors. *Cent Eur J Public Health,* 19**,** 217-21.

Azevedo, T. C., Filgueira, N. A. & Lopes, E. P. 2012. Risk factors for hepatitis C virus infection in former Brazilian soccer players. *Epidemiol Infect,* 140**,** 70-3.

Balasekaran, R., Bulterys, M., Jamal, M. M., Quinn, P. G., Johnston, D. E., Skipper, B., Chaturvedi, S. & Arora, S. 1999. A case-control study of risk factors for sporadic hepatitis C virus infection in the southwestern United States. *Am J Gastroenterol,* 94**,** 1341-6.

Bari, A., Akhtar, S., Rahbar, M. H. & Luby, S. P. 2001. Risk factors for hepatitis C virus infection in male adults in Rawalpindi-Islamabad, Pakistan. *Trop Med Int Health,* 6**,** 732-8.

Becker, T. M., Lee, F., Daling, J. R. & Nahmias, A. J. 1996. Seroprevalence of and risk factors for antibodies to herpes simplex viruses, hepatitis B, and hepatitis C among southwestern Hispanic and non-Hispanic white women. *Sex Transm Dis,* 23**,** 138-44.

Ben Alaya Bouafif, N., Triki, H., Mejri, S., Bahri, O., Chlif, S., Bettaib, J., Hechmi, S., Dellagi, K. & Ben Salah, A. 2007. A case control study to assess risk factors for hepatitis C among a general population in a highly endemic area of northwest Tunisia. *Arch Inst Pasteur Tunis,* 84**,** 21-7.

Birnie, G., Quigley, E., Clements, G., Follet, E. & Watkinson, G. 1983. Endoscopic transmission of hepatitis B virus. *Gut,* 24**,** 171-174.

Blackard, J. T., Shata, M. T., Shire, N. J. & Sherman, K. E. 2007. Acute hepatitis C virus infection: a chronic problem. *Hepatology,* 47**,** 321-331.

Bohman, V. R., Stettler, R. W., Little, B. B., Wendel, G. D., Sutor, L. J. & Cunningham, F. G. 1992. Seroprevalence and risk factors for hepatitis C virus antibody in pregnant women. *Obstet Gynecol,* 80**,** 609-13.

Bollepalli, S., Wisinger, D., Markov, M., Patel, P. & Nadir, A. 2006. Travel: A Unique Risk Factor for Acute Hepatitis C Virus. *Medscape General Medicine,* 8**,** 17.

Book, T. W. F. 2012. *Population growth rate* [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/fields/2002.html [Accessed].

Bosan, A., Qureshi, H., Bile, K. M., Ahmad, I. & Hafiz, R. 2010. A review of hepatitis viral infections in Pakistan. *J Pak Med Assoc,* 60**,** 1045-58.

Bourquin, J., Schmidli, H., Van Hoogevest, P. & Leuenberger, H. 1998. Advantages of Artificial Neural Networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. *European journal of pharmaceutical sciences,* 7**,** 5-16.

Brandao, A. B. & Fuchs, S. C. 2002. Risk factors for hepatitis C virus infection among blood donors in southern Brazil: a case-control study. *BMC Gastroenterol,* 2**,** 18.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. 1984. *Classification and regression trees*, Chapman & Hall/CRC.

Briggs, M. E., Baker, C., Hall, R., Gaziano, J. M., Gagnon, D., Bzowej, N. & Wright, T. L. 2001. Prevalence and risk factors for hepatitis C virus infection at an urban Veterans Administration medical center. *Hepatology,* 34**,** 1200-5.

Brugnano, R., Francisci, D., Quintaliani, G., Gaburri, M., Nori, G., Verdura, C., Giombini, L. & Buoncristiani, U. 1992. Antibodies against hepatitis C virus in hemodialysis patients in the central Italian region of Umbria: evaluation of some risk factors. *Nephron,* 61**,** 263-5.

Caldwell, S. H., Jeffers, L. J., Ditomaso, A., Millar, A., Clark, R. M., Rabassa, A., Reddy, K. R., De Medina, M. & Schiff, E. R. 1991. Antibody to hepatitis C is common among patients with alcoholic liver disease with and without risk factors. *Am J Gastroenterol,* 86**,** 1219-23.

Camdeviren, H., Mendes, M., Ozkan, M. M., Toros, F., Sasmaz, T. & Oner, S. 2005. Determination of depression risk factors in children and adolescents by regression tree methodology. *Acta medica Okayama,* 59**,** 19.

Camdeviren, H. A., Yazici, A. C., Akkus, Z., Bugdayci, R. & Sungur, M. A. 2007. Comparison of logistic regression model and classification tree: An application to postpartum depression data. *Expert Systems with Applications,* 32**,** 987-994.

Camilli, G. 1995. The relationship between Fisher's exact test and Pearson's chi-square test: a Bayesian perspective. *Psychometrika,* 60**,** 305-312.

Camp, N. J. & Slattery, M. L. 2002a. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes & Control,* 13**,** 813-823.

Camp, N. J. & Slattery, M. L. 2002b. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes and Control,* 13**,** 813-823.

Carney, K., Dhalla, S., Aytaman, A., Tenner, C. T. & Francois, F. 2013. Association of tattooing and hepatitis C virus infection: A multicenter case-control study. *Hepatology,* 57**,** 2117-2123.

Chang, H. C., Yu, M. W., Lu, C. F., Chiu, Y. H. & Chen, C. J. 2001. Risk factors associated with hepatitis C virus infection in Taiwanese government employees. *Epidemiol Infect,* 126**,** 291-9.

Chiaramonte, M., Stroffolini, T., Lorenzoni, U., Minniti, F., Conti, S., Floreani, A., Ntakirutimana, E., Vian, A., Ngatchu, T. & Naccarato, R. 1996. Risk factors in community-acquired chronic hepatitis C virus infection: a case-control study in Italy. *J Hepatol,* 24**,** 129-34.

Chlabicz, S., Flisiak, R., Grzeszczuk, A., Kovalchuk, O., Prokopowicz, D. & Chyczewski, L. 2006. Known and probable risk factors for hepatitis C infection: a case series in north-eastern Poland. *World J Gastroenterol,* 12**,** 141-5.

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. & Houghton, M. 1989. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science,* 244**,** 359-362.

Coates, E. A., Brennan, D., Logan, R. M., Goss, A. N., Scopacasa, B., Spencer, A. J. & Gorkic, E. 2000. Hepatitis C infection and associated oral health problems. *Australian Dental Journal,* 45**,** 108-114.

Collett, D. 2002. *Modelling binary data,* London, CHAPMAN & HALL/CRC.

Comandini, U. V., Tossini, G., Longo, M. A., Ferri, F., Cuzzi, G., Noto, P., Zaccarelli, M. & Visco, G. 1998. Sporadic hepatitis C virus infection: a case-control study of transmission routes in a selected hospital sample of the general population in Italy. *Scandinavian journal of infectious diseases,* 30**,** 11-15.

Control, U. N. O. F. D., Prevention, C. & Drugs, U. N. O. O. 2003. *Global illicit drug trends,* United Nations Publications.

Cornberg, M., Razavi, H. A., Alberti, A., Bernasconi, E., Buti, M., Cooper, C., Dalgard, O., Dillion, J. F., Flisiak, R. & Forns, X. 2011. A systematic review of hepatitis C virus epidemiology in Europe, Canada and Israel. *Liver International,* 31**,** 30-60.

Curran Jr, W. J., Scott, C. B., Horton, J., Nelson, J. S., Weinstein, A. S., Fischbach, A. J., Chang, C. H., Rotman, M., Asbell, S. O. & Krisch, R. E. 1993. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *Journal of the National Cancer Institute,* 85**,** 704-710.

Darwish, M. A., Raouf, T. A., Rushdy, P., Constantine, N. T., Rao, M. R. & Edelman, R. 1993. Risk factors associated with a high seroprevalence of hepatitis C virus infection in Egyptian blood donors. *Am J Trop Med Hyg,* 49**,** 440-7.

Davaalkham, D., Ojima, T., Nymadawa, P., Uehara, R., Watanabe, M., Oki, I. & Nakamura, Y. 2006. Prevalence and risk factors for hepatitis C virus infection in Mongolian children: Findings from a nationwide survey. *J Med Virol,* 78**,** 466-72.

Dawson, B. & Trapp, R. G. 2004. *Basic & clinical biostatistics,* Lange Medical Books/McGraw-Hill New York:.

De Waure, C., Cefalo, C., Chiaradia, G., Sferrazza, A., Miele, L., Gasbarrini, G., Ricciardi, W., Grieco, A. & La Torre, G. 2010. Intrafamilial transmission of hepatitis C virus in Italy: a systematic review. *Journal of epidemiology and community health,* 64**,** 843-848.

Delage, G., Infante-Rivard, C., Chiavetta, J. A., Willems, B., Pi, D. & Fast, M. 1999. Risk factors for acquisition of hepatitis C virus infection in blood donors: results of a case-control study. *Gastroenterology,* 116**,** 893-9.

Delarocque-Astagneau, E., Pillonel, J., De Valk, H., Perra, A., Laperche, S. & Desenclos, J. C. 2007. An incident case–control study of modes of hepatitis C virus transmission in France. *Annals of epidemiology,* 17**,** 755-762.

Dentico, P., Buongiorno, R., Volpe, A., Carlone, A., Carbone, M., Manno, C., Proscia, F., Pastore, G. & Schiraldi, O. 1992. Prevalence and incidence of hepatitis C virus (HCV) in hemodialysis patients: study of risk factors. *Clin Nephrol,* 38**,** 49-52.

Deuffic, S., Poynard, T. & Valleron, A. J. 2002. Correlation between hepatitis C virus prevalence and hepatocellular carcinoma mortality in Europe. *Journal of Viral Hepatitis,* 6**,** 411-413.

Dreiseitl, S. & Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics,* 35**,** 352-359.

Dussol, B., Verdier, J. M., Goff, J. M. L., Berthezene, P. & Berland, Y. 2007. Artificial neural networks for assessing the risk factors for urinary calcium stones according to gender and family history of stone. *Scandinavian journal of urology and nephrology,* 41**,** 414-418.

Dwyre, D., Fernando, L. & Holland, P. 2011. Hepatitis B, hepatitis C and HIV transfusion-transmitted infections in the 21st century. *Vox sanguinis,* 100**,** 92-98.

El-Sadawy, M., Ragab, H., El-Toukhy, H., El-Mor Ael, L., Mangoud, A. M., Eissa, M. H., Afefy, A. F., El-Shorbagy, E., Ibrahem, I. A., Mahrous, S., Abdel-Monem, A., Sabee, E. I., Ismail, A., Morsy, T. A., Etewa, S., Nor Edin, E., Mostafa, Y., Abouel-Magd, Y., Hassan, M. I., Lakouz, K., Abdel-Aziz, K., El-Hady, G. & Saber, M. 2004. Hepatitis C virus infection at Sharkia Governorate, Egypt: seroprevalence and associated risk factors. *J Egypt Soc Parasitol,* 34**,** 367-84.

El-Serag, H. B. 2002. Hepatocellular carcinoma: an epidemiologic view. *Journal of clinical gastroenterology,* 35**,** S72-S78.

Eusaph, A. Z., Iqbal, S., Rana, T. & Asghar, F. 2011. Evaluation of Practices of Blood Transfusion in Various Indication of Caesarean Section. *Annals of King Edward Medical University,* 17.

Falconer, J. A., Naughton, B. J., Dunlop, D. D., Roth, E. J., Strasser, D. C. & Sinacore, J. M. 1994. Predicting stroke inpatient rehabilitation outcome using a classification tree approach. *Archives of Physical Medicine and Rehabilitation,* 75**,** 619.

Field, A. 2000. *Discovering Statistics Using SPSS for Windows: Advanced Techniques for Beginners,* SAGE Publications.

Flaherty, D. C. W. & Patterson, D. A. 2003. Predicting child physical abuse recurrence: comparison of a neural network to logistic regression. *Journal of Technology in Human Services,* 21**,** 93-111.

Flamm, S. L., Parker, R. A. & Chopra, S. 1998. Risk factors associated with chronic hepatitis C virus infection: limited frequency of an unidentified source of transmission. *Am J Gastroenterol,* 93**,** 597-600.

Frank, C., Mohamed, M. K., Strickland, G. T., Lavanchy, D., Arthur, R. R., Magder, L. S., Khoby, T. E., Abdel-Wahab, Y., Ohn, E. S. A. & Anwar, W. 2000. The role of

parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *The Lancet,* 355**,** 887-891.

Fuiano, B., Pannullo, A., Annovazzi, G., D'anna, C., Materazzetti, D., Nestola, G. & Puoti, C. 1992. Risk factors and association with HBV infection in chronic C hepatitis. *Ital J Gastroenterol,* 24**,** 409-11.

Gao, W., Wang, S., Wang, Z., Shi, L. & Dong, F. 2004. Study on the application of artificial neural network in analysing the risk factors of diabetes mellitus]. *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi,* 25**,** 715.

Garassini, M. E., Pulgar, Y., Alvarado, M. & Garassini, M. A. 1995. [Hepatitis caused by virus C. Risk factors]. *G E N,* 49**,** 189-95.

Garten, R. J., Lai, S., Zhang, J., Liu, W., Chen, J., Vlahov, D. & Yu, X. F. 2004. Rapid transmission of hepatitis C virus among young injecting heroin users in Southern China. *International journal of epidemiology,* 33**,** 182-188.

Gates, J. A., Post, J. J., Kaldor, J. M., Pan, Y., Haber, P. S., Lloyd, A. R. & Dolan, K. A. 2004. Risk factors for hepatitis C infection and perception of antibody status among male prison inmates in the Hepatitis C Incidence and Transmission in Prisons Study cohort, Australia. *J Urban Health,* 81**,** 448-52.

Ghaffar, A., Mureed, S. & Chapman, R. S. 2009. RISK FACTORS FOR HEPATITIS C AMONG WOMEN OF REPRODUCTIVE AGE: A CASE CONTROL STUDY IN QUETTA, PAKISTAN. *J Health Res,* 23**,** 59-63.

Ghani, A. 2012. *National hepatitis control programme stands nowhere?* [Online]. Available: http://www.nation.com.pk/pakistan-news-newspaper-daily-english-online/islamabad/28-Jul-2012/-national-hepatitis-control-programme-stands-nowhere [Accessed].

Gheorghe, L., Csiki, I. E., Iacob, S., Gheorghe, C., Smira, G. & Regep, L. 2010. The prevalence and risk factors of hepatitis C virus infection in adult population in Romania: a nationwide survey 2006 - 2008. *J Gastrointestin Liver Dis,* 19**,** 373-9.

Ghias, M. & Pervaiz, M. K. 2009. Identification of epidemiological risk factors for hepatitis c in Punjab, Pakistan. *J Ayub Med Coll Abbottabad,* 21**,** 156-161.

Ghias, M., Pervaiz, M. K. & Aslam, A. 2010. Risk Factors for Hepatitis C Virus among Urban/Rural Settings of Patients Visiting Tertiary Care Hospitals at Lahore, Pakistan. *Journal of Statistics,* 17**,** 33-46.

Ghias, M., Pervaiz, M. K., Marshal, R. & Thornely, S. 2012a. Identification of Risk factors for hepatitis C in the Gujranwala district of Punjab, Pakistan. *World applied sciences journal,* 20**,** 94-101.

Ghias, M., Pervaiz, M. K., Thornley, S. & Marshall, R. 2012b. Statistical Modelling and Analysis of Risk Factors for Hepatitis C Infection In Punjab, Pakistan. *World applied sciences journal,* 20**,** 241-252.

Girardi, E., Zaccarelli, M., Tossini, G., Puro, V., Narciso, P. & Visco, G. 1990. Hepatitis C virus infection in intravenous drug users: prevalence and risk factors. *Scand J Infect Dis,* 22**,** 751-2.

González-Candelas, F., Guiral, S., Carbó, R., Valero, A., Vanaclocha, H., González, F. & Bracho, M. A. 2010. Patient-to-patient transmission of hepatitis C virus (HCV) during colonoscopy diagnosis. *Virol J [serie en internet],* 7**,** 217.

Govt., P. 2007. Punjab Economic Report. Available: http://www.pndpunjab.gov.pk/user_files/File/PunjabEconomicReport2007-08.pdf.

Grebely, J., Prins, M., Hellard, M., Cox, A. L., Osburn, W. O., Lauer, G., Page, K., Lloyd, A. R. & Dore, G. J. 2012. Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine. *The Lancet infectious diseases,* 12**,** 408-414.

Green, M., Björk, J., Forberg, J., Ekelund, U., Edenbrandt, L. & Ohlsson, M. 2006. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial intelligence in medicine,* 38**,** 305-318.

Guadagnino, V., Stroffolini, T., Rapicetta, M., Costantino, A., Kondili, L. A., Menniti-Ippolito, F., Caroleo, B., Costa, C., Griffo, G., Loiacono, L., Pisani, V., Foca, A. & Piazza, M. 1997. Prevalence, risk factors, and genotype distribution of hepatitis C virus infection in the general population: a community-based survey in southern Italy. *Hepatology,* 26**,** 1006-11.

Guimaraes, T., Granato, C. F., Varella, D., Ferraz, M. L., Castelo, A. & Kallas, E. G. 2001. High prevalence of hepatitis C infection in a Brazilian prison: identification of risk factors for infection. *Braz J Infect Dis,* 5**,** 111-8.

Guo, H., Shyu, Y. & Chang, H. 2006. Combining logistic regression with classification and regression tree to predict quality of care in a home health nursing data set. *Studies in health technology and informatics,* 122**,** 891.

Habib, M., Mohamed, M. K., Abdel-Aziz, F., Magder, L. S., Abdel-Hamid, M., Gamil, F., Madkour, S., Mikhail, N. N., Anwar, W., Strickland, G. T., Fix, A. D. & Sallam, I. 2001. Hepatitis C virus infection in a community in the Nile Delta: risk factors for seropositivity. *Hepatology,* 33**,** 248-53.

Hagan, M. T., Demuth, H. B. & Beale, M. H. 1996. *Neural network design,* Thomson Learning Stamford, CT.

Hajiani, E., Hashemi, J., Masjedizadeh, R., Shayesteh, A. A., Idani, E. & Rajabi, T. 2006a. Seroepidemiology of hepatitis C and its risk factors in Khuzestan Province, south-west of Iran: a case-control study. *World J Gastroenterol,* 12**,** 4884-7.

Hajiani, E., Masjedizadeh, R., Hashemi, J., Azmi, M. & Rajabi, T. 2006b. Hepatis C virus transmission and its risk factors within families of patients infected with hepatitis C virus in southern Iran: Khuzestan. *World J Gastroenterol,* 12**,** 7025-8.

Hakimpoor, H., Arshad, K. A. B., Tat, H. H., Khani, N. & Rahmandoust, M. 2011. Artificial Neural Networks' Applications in Management. *World Applied Sciences Journal,* 14**,** 1008-1019.

Halim, N. K. & Ajayi, O. I. 2000. Risk factors and seroprevalence of hepatitis C antibody in blood donors in Nigeria. *East Afr Med J,* 77**,** 410-2.

Hall, G. F. 2007. Hepatitis A, B, C, D, E, G: an update. *Ethn Dis,* 17**,** S2-40.

Hand, W. L. & Vasquez, Y. 2005. Risk factors for hepatitis C on the Texas-Mexico border. *Am J Gastroenterol,* 100**,** 2180-5.

Hart, A. & Wyatt, J. 1990. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Informatics for Health and Social Care,* 15**,** 229-236.

Hasford, J., Ansari, H. & Lehmann, K. 1993. CART and logistic regression analyses of risk factors for first dose hypotension by an ACE-inhibitor. *Therapie,* 48**,** 479.

Hashmi, A., Saleem, K. & Soomro, J. A. 2010. Prevalence and factors associated with hepatitis C virus seropositivity in female individuals in islamabad, pakistan. *International journal of preventive medicine,* 1**,** 252.

Hatcher, L. & Institute, S. 2003. *Step-by-step Basic Statistics Using Sas: Student Guide,* Sas Inst.

He, Y., Zhang, J., Zhong, L., Chen, X., Liu, H. M., Wan, L. K., Wang, H., Li, H., Tian, L., Hu, J. L., Luo, P., Wang, L., Chen, Y., Liu, T., Liu, S. L. & Lu, W. B. 2011. Prevalence of and risk factors for hepatitis C virus infection among blood donors in Chengdu, China. *J Med Virol,* 83**,** 616-21.

Hellard, M., Sacks-Davis, R. & Gold, J. 2009. Hepatitis C treatment for injection drug users: a review of the available evidence. *Clinical Infectious Diseases,* 49**,** 561-573.

Hepburn, M. J. & Lawitz, E. J. 2004. Seroprevalence of hepatitis C and associated risk factors among an urban population in Haiti. *BMC Gastroenterol,* 4**,** 31.

Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N. & Abbruzzese, J. L. 1999. Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clinical cancer research,* 5**,** 3403-3410.

Heydari, S. T., Ayatollahi, S. M. T. & Zare, N. 2011. Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity. *Journal of Medical Systems***,** 1-6.

Hies. 2011. *Household Integrated Economic Survey (HIES) 2010-11* [Online]. Islamabad: Pakistan Bureau of Statistics. Available: http://www.pbs.gov.pk/content/household-integrated-economic-survey-hies-2010-11 [Accessed 1st January, 2014].

Ho, E., Ha, N., Ahmed, A., Ayoub, W., Daugherty, T., Garcia, G., Cooper, A., Keeffe, E. & Nguyen, M. 2012a. Prospective study of risk factors for hepatitis C virus acquisition by Caucasian, Hispanic, and Asian American patients. *Journal of Viral Hepatitis*.

Ho, E. Y., Ha, N. B., Ahmed, A., Ayoub, W., Daugherty, T., Garcia, G., Cooper, A., Keeffe, E. B. & Nguyen, M. H. 2012b. Prospective study of risk factors for hepatitis C virus acquisition by Caucasian, Hispanic, and Asian American patients. *J Viral Hepat,* 19**,** e105-11.

Holmberg, S. 2012. *CDC Health Information for Inter-national Travel 2012 (yellow book),* UK, Oxford University Press.

Hong, W., Ji, Y., Wang, D., Chen, T. & Zhu, Q. 2011. Use of artificial neural network to predict esophageal varices in patients with HBV related cirrhosis. *Hepatitis Monthly,* 11**,** 544.

Hosmer, D. W. & Lemeshow, S. 2000. *Applied logistic regression,* Canada, Wiley-Interscience.

Hosseini-Moghaddam, S. M., Keyvani, H., Kasiri, H., Kazemeyni, S. M., Basiri, A., Aghel, N. & Alavian, S. M. 2006. Distribution of hepatitis C virus genotypes among hemodialysis patients in Tehran—a multicenter study. *Journal of medical virology,* 78**,** 569-573.

Howell, D. C. 2009. Chi-Square Test Analysisof Contigency Tables *Women,* 35**,** 1-4.

Idrees, M., Lal, A., Naseem, M. & Khalid, M. 2008. High prevalence of hepatitis C virus infection in the largest province of Pakistan. *Journal of digestive diseases,* 9**,** 95-103.

Idrees, M. & Riazuddin, S. 2008. Frequency distribution of hepatitis C virus genotypes in different geographical regions of Pakistan and their possible routes of transmission. *BMC Infectious Diseases,* 8**,** 69.

Ijaz, R. M. & Akhter, A. S. 2007. Evaluation of Risk Factors of HCV infection in Lahore,(Pakistan). *Pakistan Journal of Statistics and Operation Research,* 3.

Iqbal, Z., Haque, A. R., Keynan, M. H. A., Farah, I., Mukhtar, E. M., Aijaz, S., Niazi, H. K., Ahmed, M. S., Fakhar-Ud-Din, M. & Khan, M. A. S. 2003. Description and Analysis.

Irfan, A. & Arfeen, S. 2004. Hepatitis C virus infection in spouses. *Pak J Med Res,* 43**,** 113-6.

Irfan, M., Nadeem, M. A., Mirza, H. G., Ghias, M., Mohsin, A. & Muttee, M. U. K. 2011. Statistical prediction model for relapse rate in chronic hepatitis C patients treated with conventional interferon and ribavirin therapy. *British Journal of Medicine and Medical Research,* 1**,** 122-131.

Jafri, W., Jafri, N., Yakoob, J., Islam, M., Tirmizi, S. F., Jafar, T., Akhtar, S., Hamid, S., Shah, H. A. & Nizami, S. Q. 2006. Hepatitis B and C: prevalence and risk factors associated with seropositivity among children in Karachi, Pakistan. *BMC Infect Dis,* 6**,** 101.

Jafri, W. & Subhan, A. 2010. Hepatitis C in Pakistan: magnitude, genotype, disease characteristics and therapeutic response. *Tropical Gastroenterology,* 29**,** 194-201.

Jajoo, R., Mital, D., Haque, S. & Srinivasan, S. 2002. Prediction of hepatitis C using artificial neural network. 3**,** 1545-1550

Jamil, M. S., Ali, H., Shaheen, R. & Basit, A. 2010. PREVALENCE, KNOWLEDGE AND AWARENESS OF HEPATITIS C AMONG RESIDENTS OF THREE UNION COUNCILS IN MANSEHRA. *J Ayub Med Coll Abbottabad,* 22.

Janjua, N. & Nizamy, M. 2004. Knowledge and practices of barbers about hepatitis B and C transmission in Rawalpindi and Islamabad. *JOURNAL-PAKISTAN MEDICAL ASSOCIATION,* 54**,** 116-118.

Janjua, N. Z., Akhtar, S. & Hutin, Y. J. F. 2005. Injection use in two districts of Pakistan: implications for disease prevention. *International Journal for Quality in Health Care,* 17**,** 401-408.

Janjua, N. Z., Hamza, H. B., Islam, M., Tirmizi, S. F., Siddiqui, A., Jafri, W. & Hamid, S. 2010. Health care risk factors among women and personal behaviours among men explain the high prevalence of hepatitis C virus infection in Karachi, Pakistan. *J Viral Hepat,* 17**,** 317-26.

Janjua, N. Z., Hutin, Y. J., Akhtar, S. & Ahmad, K. 2006a. Population beliefs about the efficacy of injections in Pakistan's Sindh province. *Public Health,* 120**,** 824-833.

Janjua, N. Z., Khan, A. J., Altaf, A. & Ahmad, K. 2006b. Towards Safe Injection Practices in Pakistan.

Jatapai, A., Nelson, K. E., Chuenchitra, T., Kana, K., Eiumtrakul, S., Sunantarod, E. & Rangsin, R. 2010. Prevalence and risk factors for hepatitis C virus infection among young Thai men. *Am J Trop Med Hyg,* 83**,** 433-9.

Jatoi, S. M., Narsani, A. K. & Kumar, M. 2009. Frequency of Anti Hepatitis C Virus in Eye Surgery Patients at Tertiary Referral Center LUMHS. *Pak J Ophthalmol,* 25.

Jl, F. 1981. *Statistical Methods for Rates and Proportions,* John Wiley & Sons.

Kahn, H. A. & Sempos, C. T. 1989. *Statistical Methods in Epidemiology,* Oxford University Press.

Kaldor, J. M., Archer, G. T., Buring, M. L., Ismay, S. L., Kenrick, K. G., Lien, A. S., Purusothaman, K., Tulloch, R., Bolton, W. V. & Wylie, B. R. 1992. Risk factors for hepatitis C virus infection in blood donors: a case-control study. *Med J Aust,* 157**,** 227-30.

Kandeel, A., Talaat, M., Afifi, S., El-Sayed, N., Fadeel, M. A., Hajjeh, R. & Mahoney, F. 2012. Case control study to identify risk factors for acute hepatitis C virus infection in Egypt. *BMC Infectious Diseases,* 12**,** 294.

Kao, J.-H., Chen, P.-J., Yang, P.-M., Lai, M.-Y., Sheu, J.-C., Wang, T.-H. & Chen, D.-S. 1992. Intrafamilial transmission of hepatitis C virus: the important role of infections between spouses. *Journal of Infectious Diseases,* 166**,** 900-903.

Karaca, C., Cakaloglu, Y., Demir, K., Ozdil, S., Kaymakoglu, S., Badur, S. & Okten, A. 2006. Risk factors for the transmission of hepatitis C virus infection in the Turkish population. *Dig Dis Sci,* 51**,** 365-9.

Karim, F., Foster, G., Akbar, S. & Rahman, S. 2011. Prevalence and Risk Factors of Asymptomatic Hepatitis C Virus Infection in Bangladesh. *Journal of Clinical and Experimental Hepatology,* 1**,** 13-16.

Karmochkine, M., Carrat, F., Dos Santos, O., Cacoub, P. & Raguin, G. 2006a. A case-control study of risk factors for hepatitis C infection in patients with unexplained routes of infection. *J Viral Hepat,* 13**,** 775-82.

Karmochkine, M., Carrat, F., Dos Santos, O., Cacoub, P. & Raguin, G. 2006b. A case–control study of risk factors for hepatitis C infection in patients with unexplained routes of infection*. *Journal of Viral Hepatitis,* 13**,** 775-782.

Kasiulevičius, V., Šapoka, V. & Filipavičiūtė, R. 2006. Sample size calculation in epidemiological studies. *Gerantologija,* 7**,** 225-231.

Kazemnejad, A., Batvandi, Z. & Faradmal, J. 2010. Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes. *East Mediterr Health J,* 16**,** 615-620.

Kermode, M. 2004. Unsafe injections in low-income country health settings: need for injection safety promotion to prevent the spread of blood-borne viruses. *Health promotion international,* 19**,** 95-103.

Kerzman, H., Green, M. S. & Shinar, E. 2007. Risk factors for hepatitis C virus infection among blood donors in Israel: a case-control study between native Israelis and immigrants from the former Soviet Union. *Transfusion,* 47**,** 1189-96.

Khan, F. A., Khan, M., Ali, A. & Chohan, U. 2006. Estimation of blood loss during Caesarean Section: an audit. *Pakistan journal of medical association,* 56**,** 572.

Khan, U. R., Janjua, N. Z., Akhtar, S. & Hatcher, J. 2008a. Case-control study of risk factors associated with hepatitis C virus infection among pregnant women in hospitals of Karachi-Pakistan. *Trop Med Int Health,* 13**,** 754-61.

Khan, U. R., Janjua, N. Z., Akhtar, S. & Hatcher, J. 2008b. Case–control study of risk factors associated with hepatitis C virus infection among pregnant women in hospitals of Karachi-Pakistan. *Tropical Medicine & International Health,* 13**,** 754-761.

Khuwaja, A., Qureshf, F. & Fatmi, Z. 2002. Knowledge about hepatitis B and C among patients attending family. *Eastern Mediterranean health journal,* 8**,** 787.

Kim, Y. S., Ahn, Y. O. & Kim, D. W. 1996. A case-control study on the risk factors of hepatitis C virus infection among Koreans. *J Korean Med Sci,* 11**,** 38-43.

Kim, Y. S., Ahn, Y. O. & Lee, H. S. 2002. Risk factors for hepatitis C virus infection among Koreans according to the hepatitis C virus genotype. *J Korean Med Sci,* 17**,** 187-92.

Kirkwood, B. & Sterne, J. 2003. *Essential Medical Statistics*, John Wiley & Sons.

Kitsantas, P., Hollander, M. & Li, L. 2006. Using classification trees to assess low birth weight outcomes. *Artificial intelligence in medicine,* 38**,** 275-289.

Kleinbaum, D. G., Klein, M. & Pryor, E. R. 2002. *Logistic Regression: A Self-Learning Text,* Springer-Verlag.

Kolho, E. K. & Krusius, T. 1992. Risk factors for hepatitis C virus antibody positivity in blood donors in a low-risk country. *Vox Sang,* 63**,** 192-7.

Kuchibhatla, M. & Fillenbaum, G. G. 2002. Assessing risk factors for mortality in elderly White and African American people: implications of alternative analyses. *The Gerontologist,* 42**,** 826-834.

Lago, C. 2011. *The Handbook of Transcultural Counselling and Psychotherapy*, McGraw-Hill Education.

Landau, S. & Everitt, B. S. 2004. *A Handbook of Statistical Analyses Using Spss*, Chapman and Hall.

Larasati, A., Deyong, C. & Slevitch, L. Comparing Neural Network and Ordinal Logistic Regression to Analyze Attitude Responses. *Service Science***,** 304.

Lasher, L. E., Elm, J. L., Hoang, Q., Nekomoto, T. S., Cashman, T. M., Miller, F. D. & Effler, P. V. 2005. A case control investigation of hepatitis C risk factors in Hawaii. *Hawaii Med J,* 64**,** 296-300, 302-4.

Lauer, G. M. & Walker, B. D. 2001. Hepatitis C virus infection. *New England journal of medicine,* 345**,** 41-52.

Lavanchy, D. 2009. The global burden of hepatitis C. *Liver International,* 29**,** 74-81.

Lee, M. H., Yang, H. I., Jen, C. L., Lu, S. N., Yeh, S. H., Liu, C. J., You, S. L., Sun, C. A., Wang, L. Y., Chen, W. J. & Chen, C. J. 2011. Community and personal risk factors for hepatitis C virus infection: a survey of 23,820 residents in Taiwan in 1991-2. *Gut,* 60**,** 688-94.

Leke-Betechuoh, B., Marwala, T., Tim, T. & Lagazio, M. Year. Prediction of HIV Status from Demographic Data Using Neural Networks. *In*, 2006. IEEE, 2339-2344.

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. & Rakowski, W. 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine,* 26**,** 172-181.

Lewallen, S. & Courtright, P. 1998. Epidemiology in practice: case-control studies. *Community Eye Health,* 11**,** 57.

Li, C.-P., Zhi, X.-Y., Ma, J., Cui, Z., Zhu, Z.-L., Zhang, C. & Hu, L.-P. 2012. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chinese Medical Journal,* 125**,** 851-857.

Liew, P. L., Lee, Y. C., Lin, Y. C., Lee, T. S., Lee, W. J., Wang, W. & Chien, C. W. 2007. Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients. *Digestive and liver disease,* 39**,** 356-362.

Lin, C. C., Hwang, S. J., Chiou, S. T., Kuan, C. L., Chen, L. W., Lee, T. C., Lee, M. B., Lee, H. H., Hsu, P. S. & Tsai, S. T. 2003. The prevalence and risk factors analysis of serum antibody to hepatitis C virus in the elders in northeast Taiwan. *J Chin Med Assoc,* 66**,** 103-8.

Linder, R., König, I., Weimar, C., Diener, H., Pöppl, S. & Ziegler, A. 2006. A Comparison of Logistic Regression and Neural Networks. *Methods Inf Med,* 45**,** 536-40.

Lippmann, R. 1987. An introduction to computing with neural nets. *ASSP Magazine, IEEE,* 4**,** 4-22.

Liu, F., Chen, K., He, Z., Ning, T., Pan, Y., Cai, H. & Ke, Y. 2009. Hepatitis C seroprevalence and associated risk factors, Anyang, China. *Emerg Infect Dis,* 15**,** 1819-22.

Luby, S., Khanani, R., Zia, M., Vellani, Z., Ali, M., Qureshi, A., Khan, A., Mujeeb, S., Shah, S. & Fisher-Hoch, S. 2006. Evaluation of blood bank practices in Karachi, Pakistan, and the government's response. *JPMA. The Journal of the Pakistan Medical Association,* 56**,** S25.

Luby, S., Qamruddin, K., Shah, A., Omair, A., Pahsa, O., Khan, A., Mccormick, J., Hoodbhouy, F. & Fisher-Hoch, S. 1997. The relationship between therapeutic injections and high prevalence of hepatitis C infection in Hafizabad, Pakistan. *Epidemiology and infection,* 119**,** 349-356.

Luksamijarulkul, P. & Deangbubpha, A. 1997. Hepatitis C antibody prevalence and risk factors of some female sex workers in Thailand. *Southeast Asian J Trop Med Public Health,* 28**,** 507-12.

Madhava, V., Burgess, C. & Drucker, E. 2002. Epidemiology of chronic hepatitis C virus infection in sub-Saharan Africa. *The Lancet infectious diseases,* 2**,** 293-302.

Madurga Revilla, P., Aguar Carrascosa, M., Pereda Perez, A., Modesto Alapont, V., Montanes Sanchez, A., Torres Martinez, E., Brugada Montanter, M. & Leon Carinena, S. 2012. [Retrospective study of risk factors of vertical transmission of hepatitis C virus.]. *An Pediatr (Barc)*.

Magder, L. S., Fix, A. D., Mikhail, N. N. H., Mohamed, M. K., Abdel-Hamid, M., Abdel-Aziz, F., Medhat, A. & Strickland, G. T. 2005. Estimation of the risk of transmission of hepatitis C between spouses in Egypt based on seroprevalence data. *International journal of epidemiology,* 34**,** 160-165.

Maggi, G., Armitano, S., Brambilla, L., Brenna, M., Cairo, M., Galvani, G., Gola, D., Komla-Ebri, K., Marmondi, E., Perricone, G., Posca, M., Vegezzi, P. G., Vergani, C. & De Leo, G. 1999. Hepatitis C infection in an Italian population not selected for risk factors. *Liver,* 19**,** 427-31.

Maher, L., Chant, K., Jalaludin, B. & Sargent, P. 2004. Risk behaviors and antibody hepatitis B and C prevalence among injecting drug users in south-western Sydney, Australia. *Journal of gastroenterology and hepatology,* 19**,** 1114-1120.

Maimon, O. Z. & Rokach, L. 2005. *Data Mining And Knowledge Discovery Handbook*, Springer Science+Business Media.

Maisonneuve, P., Aymard, J. P., Lemaire, J. M., Baillet, A. & Janot, C. 1991. [Serum antibodies to hepatitis C virus: analysis of risk factors of seropositivity in a population of blood donors in metropolitan France]. *Rev Med Interne,* 12**,** 416-8.

Mast, E. E., Alter, M. J. & Margolis, H. S. 1999. Strategies to prevent and control hepatitis B and C virus infections: a global perspective. *Vaccine,* 17**,** 1730-1733.

Matsui, Y., Egawa, S., Tsukayama, C., Terai, A., Kuwao, S., Baba, S. & Arai, Y. 2002. Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the Japanese population. *Japanese journal of clinical oncology,* 32**,** 530-535.

Medhat, A., Shehata, M., Magder, L. S., Mikhail, N., Abdel-Baki, L., Nafeh, M., Abdel-Hamid, M., Strickland, G. T. & Fix, A. D. 2002. Hepatitis c in a community in Upper Egypt: risk factors for infection. *Am J Trop Med Hyg,* 66**,** 633-8.

Mele, A., Sagliocca, L., Manzillo, G., Converti, F., Amoroso, P., Stazi, M. A., Ferrigno, L., Rapicetta, M., Franco, E., Adamo, B. & Et Al. 1994. Risk factors for acute non-A, non-B hepatitis and their relationship to antibodies for hepatitis C virus: a case-control study. *Am J Public Health,* 84**,** 1640-3.

Menard, S. 2002. *Applied Logistic Regression Analysis*, SAGE Publications.

Mendes-Correa, M. C., Barone, A. A. & Gianini, R. J. 2005. Risk factors associated with hepatitis C among patients co-infected with human immunodeficiency virus: a case-control study. *Am J Trop Med Hyg,* 72**,** 762-7.

Mendes-Correa, M. C., Barone, A. A. & Guastini, C. 2001. Hepatitis C virus seroprevalence and risk factors among patients with HIV infection. *Rev Inst Med Trop Sao Paulo,* 43**,** 15-9.

Merle, V., Goria, O., Gourier-Frery, C., Benguigui, C., Michel, P., Huet, P., Czernichow, P. & Colin, R. 1999. [Risk factors of contamination by hepatitis C virus. Case-control study in the general population]. *Gastroenterol Clin Biol,* 23**,** 439-46.

Mics. 2011. Multiple Indicator Cluster Survey (MICS), Punjab. 1. Available: http://www.bos.gop.pk/?q=mics2011reports.

Mikhail, N. N., Lewis, D. L., Omar, N., Taha, H., El-Badawy, A., Abdel-Mawgoud, N., Abdel-Hamid, M. & Strickland, G. T. 2007. Prospective study of cross-infection from upper-GI endoscopy in a hepatitis C-prevalent population. *Gastrointestinal Endoscopy,* 65**,** 584-588.

Miranda, A. E., Figueiredo, N. C., Schmidt, R. & Page-Shafer, K. 2008. A population-based survey of the prevalence of HIV, syphilis, hepatitis B and hepatitis C infections, and associated risk factors among young women in Vitoria, Brazil. *AIDS Behav,* 12**,** S25-31.

Mishra, G., Sninsky, C., Roswell, R., Fitzwilliam, S. & Hyams, K. C. 2003. Risk factors for hepatitis C virus infection among patients receiving health care in a Department of Veterans Affairs hospital. *Dig Dis Sci,* 48**,** 815-20.

Mohamed, M. K., Abdel-Hamid, M., Mikhail, N. N., Abdel-Aziz, F., Medhat, A., Magder, L. S., Fix, A. D. & Strickland, G. T. 2005. Intrafamilial transmission of hepatitis C in Egypt. *Hepatology,* 42**,** 683-687.

Mohamed, M. K., Hussein, M. H., Massoud, A. A., Rakhaa, M. M., Shoeir, S., Aoun, A. A. & Aboul Naser, M. 1996. Study of the risk factors for viral hepatitis C infection among Egyptians applying for work abroad. *J Egypt Public Health Assoc,* 71**,** 113-47.

Mohamed, N., Ahmad, W. M. A. W., Aleng, N. A. & Ahmad, M. H. 2011. Assessing the efficiency of multilayer feed-forward neural network model: Application to body mass index data. *World Applied Sciences Journal,* 15**,** 677-682.

Mohamoud, Y. A., Mumtaz, G. R., Riome, S., Miller, D. & Abu-Raddad, L. J. 2013. The epidemiology of hepatitis C virus in Egypt: a systematic review and data synthesis. *BMC Infectious Diseases,* 13**,** 288.

Molenberghs, G. 2003. Statistical methodology in biometry. *Biometrics,* 1**,** 1-20.

Møller, M. F. 1993. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks,* 6**,** 525-533.

Montagnac, R., Schillinger, F., Finger, L., Abid, A. & Oudin, M. 1994. [Hepatitis C in hemodialysis: contribution of a serum bank for the evaluation of risk factors]. *Presse Med,* 23**,** 181.

Moriya, T., Sasaki, F., Mizui, M., Ohno, N., Mohri, H., Mishiro, S. & Yoshizawa, H. 1995. Transmission of hepatitis C virus from mothers to infants: its frequency and risk factors revisited. *Biomed Pharmacother,* 49**,** 59-64.

Mostafa, A., Taylor, S. M., El-Daly, M., El-Hoseiny, M., Bakr, I., Arafa, N., Thiers, V., Rimlinger, F., Abdel-Hamid, M., Fontanet, A. & Mohamed, M. K. 2010. Is the hepatitis C virus epidemic over in Egypt? Incidence and risk factors of new hepatitis C virus infections. *Liver Int,* 30**,** 560-6.

Muhammad, N. & Jan, M. A. 2005. Frequency of hepatitis" C" in Buner, NWFP. *Journal of the College of Physicians and Surgeons--Pakistan: JCPSP,* 15**,** 11.

Murphy, E. L., Bryzman, S. M., Glynn, S. A., Ameti, D. I., Thomson, R. A., Williams, A. E., Nass, C. C., Ownby, H. E., Schreiber, G. B. & Kong, F. 2000a. Risk factors for hepatitis C virus infection in United States blood donors. *Hepatology,* 31**,** 756-762.

Murphy, E. L., Bryzman, S. M., Glynn, S. A., Ameti, D. I., Thomson, R. A., Williams, A. E., Nass, C. C., Ownby, H. E., Schreiber, G. B., Kong, F., Neal, K. R. & Nemo, G. J. 2000b. Risk factors for hepatitis C virus infection in United States blood donors. NHLBI Retrovirus Epidemiology Donor Study (REDS). *Hepatology,* 31**,** 756-62.

Mustufa, M. A., Memon, A. A., Nasim, S., Shahid, A. & Omar, S. M. 2010. Exposure to risk factors for hepatitis B and C viruses among primary school teachers in Karachi. *J Infect Dev Ctries,* 4**,** 616-20.

Muzaffar, F., Hussain, I. & Haroon, T. S. 2008. Hepatitis C: the dermatologic profile. *J Pak Assoc Dermatol,* 18**,** 171-181.

Nagy, K., Reiczigel, J., Harnos, A., Schrott, A. & Kabai, P. 2010. Tree-based methods as an alternative to logistic regression in revealing risk factors of crib-biting in horses. *Journal of Equine Veterinary Science,* 30**,** 21-26.

Nakashima, K., Kashiwagi, S., Hayashi, J., Noguchi, A., Hirata, M., Ikeda, S., Sakota, I. & Shingu, T. 1993. Low prevalence of hepatitis C virus infection among hospital staff and acupuncturists in Kyushu, Japan. *Journal of Infection,* 26**,** 17-25.

Ndong-Atome, G. R., Njouom, R., Padilla, C., Bisvigou, U., Makuwa, M. & Kazanji, M. 2009. Absence of intrafamilial transmission of hepatitis C virus and low risk for sexual transmission in rural central Africa indicate a cohort effect. *Journal of Clinical Virology,* 45**,** 349-353.

Neal, K. R., Jones, D. A., Killey, D. & James, V. 1994. Risk factors for hepatitis C virus infection. A case-control study of blood donors in the Trent Region (UK). *Epidemiol Infect,* 112**,** 595-601.

Neale, J. & Stevenson, C. 2012. Routine exposure to blood within hostel environments might help to explain elevated levels of hepatitis C amongst homeless drug users: Insights from a qualitative study. *International Journal of Drug Policy*.

Nelder, J. A. & Wedderburn, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)***,** 370-384.

Nelson, L. M., Bloch, D. A., Longstreth Jr, W. & Shi, H. 1998. Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. *Journal of clinical epidemiology,* 51**,** 199-209.

Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J. & Perelson, A. S. 1998. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy. *Science,* 282**,** 103-107.

Nguyen, V. T., Mclaws, M. L. & Dore, G. J. 2007. Prevalence and risk factors for hepatitis C infection in rural north Vietnam. *Hepatol Int,* 1**,** 387-93.

Nokhodian, Z., Meshkati, M., Adibi, P., Ataei, B., Kassaian, N., Yaran, M., Shoaei, P. & Hassannejad, R. 2012. Hepatitis C among Intravenous Drug Users in Isfahan, Iran: a Study of Seroprevalence and Risk Factors. *International journal of preventive medicine,* 3**,** S131.

Nurunnabi, A. & West, G. Year. Outlier Detection in Logistic Regression: A Quest for Reliable Knowledge from Predictive Modeling and Classification. *In:* Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, 2012. IEEE, 643-652.

Nyamathi, A. M., Dixon, E. L., Robbins, W., Smith, C., Wiley, D., Leake, B., Longshore, D. & Gelberg, L. 2002. Risk factors for hepatitis C virus infection among homeless adults. *Journal of general internal medicine,* 17**,** 134-143.

Obienu, O., Nwokediuko, S., Malu, A. & Lesi, O. A. 2011. Risk factors for hepatitis C virus transmission obscure in nigerian patients. *Gastroenterol Res Pract,* 2011**,** 939673.

Ohto, H., Terazawa, S., Sasaki, N., Hino, K., Ishiwata, C., Kako, M., Ujiie, N., Endo, C. & Matsui, A. 1994. Transmission of hepatitis C virus from mothers to infants. *New England Journal of Medicine,* 330**,** 744-750.

Oliveira-Filho, A. B. D., Pimenta, A. D. S. C., Rojas, M. D. F. M., Chagas, M. C. M., Crespo, D. M., Crescente, J. Â. B. & Lemos, J. A. R. D. 2010. Likely transmission of hepatitis C virus through sharing of cutting and perforating instruments in blood donors in the State of Pará, Northern Brazil. *Cadernos de Saúde Pública,* 26**,** 837-844.

Onyango, D., Kikuvi, G., Amukoye, E. & Omolo, J. 2013. Risk factors of severe pneumonia among children aged 2-59 months in western Kenya: a case control study. *Pan African Medical Journal,* 13.

Open-Epi. *Open Source Epidemiologic Statistics for Public Health* [Online]. Available: http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm [Accessed 14th December 2012].

Ottenbacher, K. J., Linn, R. T., Smith, P. M., Illig, S. B., Mancuso, M. & Granger, C. V. 2004. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. *Annals of epidemiology,* 14**,** 551-559.

Page, K., Hahn, J. A., Evans, J., Shiboski, S., Lum, P., Delwart, E., Tobler, L., Andrews, W., Avanesyan, L. & Cooper, S. 2009. Acute hepatitis C virus infection in young adult injection drug users: a prospective study of incident infection, resolution, and reinfection. *Journal of Infectious Diseases,* 200**,** 1216-1226.

Pak-Info. Bureau of Statistics, Punjab. Available: http://bos.gop.pk/?q=pakinfo [Accessed 22 December, 2013].

Panigrahi, A. K., Panda, S. K., Dixit, R. K., Rao, K. V., Acharya, S. K., Dasarathy, S. & Nanu, A. 1997. Magnitude of hepatitis C virus infection in India: prevalence in healthy blood donors, acute and chronic liver diseases. *Journal of medical virology,* 51**,** 167-174.

Peat, J. & Barton, B. 2008. *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*, Wiley.

Peopledaily. Pakistan asks Afghans to go back or shift to camp. Available: http://english.peopledaily.com.cn/200605/09/eng20060509_263904.html [Accessed 25th December, 2013].

Pérez, C. M., Suárez, E., Torres, E. A., Román, K. & Colón, V. 2005. Seroprevalence of hepatitis C virus and associated risk behaviours: a population-based study in San Juan, Puerto Rico. *International journal of epidemiology,* 34**,** 593-599.

Perz, J. F., Armstrong, G. L., Farrington, L. A., Hutin, Y. J. F. & Bell, B. P. 2006. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of hepatology,* 45**,** 529-538.

Petrie, A. & Sabin, C. 2009. *Medical Statistics at a Glance*, Wiley.

Petrosillo, N., Puro, V., Ippolito, G., Di Nardo, V., Albertoni, F., Chiaretti, B., Rava, L., Sommella, L., Ricci, C., Zullo, G. & Et Al. 1995. Hepatitis B virus, hepatitis C virus and human immunodeficiency virus infection in health care workers: a multiple regression analysis of risk factors. *J Hosp Infect,* 30**,** 273-81.

Piper, M. E., Loh, W. Y., Smith, S. S., Japuntich, S. J. & Baker, T. B. 2011. Using decision tree analysis to identify risk factors for relapse to smoking. *Substance use & misuse,* 46**,** 492-510.

Pmrc. *National Survey on Prevalence of Hepatitis B & C in General Population of Pakistan* [Online]. Pakistan Medical Research Council. Available: http://www.pmrc.org.pk/part-1.pdf [Accessed 6 July 2012].

Pohjanpelto, P. 1992. Risk factors connected with hepatitis C infections in Finland. *Scand J Infect Dis,* 24**,** 251-2.

Poynard, T., Yuen, M. F., Ratziu, V. & Lung Lai, C. 2003. Viral hepatitis C. *Lancet,* 362**,** 2095-2100.

Pregibon, D. 1981. Logistic regression diagnostics. *The Annals of Statistics***,** 705-724.

Qazi, M. A., Imran, A., Chuadhary, M. G. & 2011. Risk Factors For The Spread Of Hepatitis C Virus (Hcv) Infection In Bahawal Pur (South Punjab, Pakistan). *Pakistan journal of Gestroentrology,* 25.

Qazi Masroor Ali, A. I., Ghulam Muhyuddin Chuadhary, & Ijaz Ahmed Shah, N. S. 2011. Risk Factors For The Spread Of Hepatitis C Virus (Hcv) Infection In Bahawal Pur (South Punjab, Pakistan). *Pakistan journal of Gestroentrology,* 25.

Qidwai, W., Fahim, A. & Waheed, S. 2010. Hepatitis C in Pakistan-A neglected challenge. *International Journal of Hepatology,* 1**,** 5-7.

Qin, L., Yang, S., Dore, K. & Pollari, F. 2005. Modelling and Analysis of Salmonella Typhimurium Infections using Logistic Regression and Neural Network Models. 3**,** 1749-1754.

Qureshi, H., Arif, A., Ahmed, W. & Alam, S. E. 2007. HCV exposure in spouses of the index cases. *JPMA. The Journal of the Pakistan Medical Association,* 57**,** 175.

Qureshi, H., Arif, A., Riaz, K., Alam, S. E., Ahmed, W. & Mujeeb, S. A. 2009. Determination of risk factors for hepatitis B and C in male patients suffering from chronic hepatitis. *BMC Res Notes,* 2**,** 212.

Raghavendra, B. & Srivatsa, S. 2011. Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets. *International Journal of Computer Science and Security (IJCSS),* 5**,** 503.

Rahman, A. U. & Uddin, S. 2009. Statistical analysis of different socio economic factors affecting education of NW. FP (Pakistan). *Journal of applied quantitative methods,* 4**,** 88-94.

Rastegari, A., Haghdoost, A. A. & Baneshi, M. R. 2013. Factors Influencing Drug Injection History among Prisoners: A Comparison between Classification and Regression Trees and Logistic Regression Analysis. *Addiction and Health,* 5.

Redondo, M. F. & Espinosa, C. H. Year. Input selection by Multilayer Feedforward trained networks. *In*, 1999. IEEE, 1834-1839 vol. 3.

Röhrig, B., Du Prel, J.-B. & Blettner, M. 2009. Study design in medical research: part 2 of a series on the evaluation of scientific publications. *Deutsches Arzteblatt International,* 106**,** 184.

Romero-Figueroa, S., Ceballos-Salgado, E., Santillan-Arreygue, L., Miranda-Garcia, M., Rubio-Lezama, M. & Garduno-Garcia, J. J. 2012. Risk factors associated with hepatitis C virus infection in an urban population of the State of Mexico. *Arch Virol,* 157**,** 329-32.

Rossi, R. J. 2009. *Applied Biostatistics for the Health Sciences*, Wiley.

Rouse, D. J., Macpherson, C., Landon, M., Varner, M. W., Leveno, K. J., Moawad, A. H., Spong, C. Y., Caritis, S. N., Meis, P. J. & Wapner, R. J. 2006. Blood transfusion and cesarean delivery. *Obstetrics & Gynecology,* 108**,** 891.

Samarasinghe, S. 2006. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*, Taylor & Francis.

Sandhu, J., Preiksaitis, J. K., Campbell, P. M., Carriere, K. C. & Hessel, P. A. 1999. Hepatitis C prevalence and risk factors in the northern Alberta dialysis population. *Am J Epidemiol,* 150**,** 58-66.

Sarle, W. S. 2000. *How to measure importance of inputs?* [Online]. SAS instutue Inc. Available: ftp://ftp.sas.com/pub/neural/importance.html [Accessed August 29, 2012 2012].

Satti, R., Mustafa, F., Khan, M. I., Haq, T. S., Khan, Z. U., Zubair, M., Ur Rasool, S. T., Azam, M., Ajmal, M. & Qamar, R. 2012. Prevalence of Hepatitis C Virus in Urban Ghettos of Twin Cities. *Pakistan J. Zool,* 44**,** 937-943.

Schneeberger, P. M., Vos, J. & Van Dijk, W. C. 1993. Prevalence of antibodies to hepatitis C virus in a Dutch group of haemodialysis patients related to risk factors. *J Hosp Infect,* 25**,** 265-70.

Serfaty, L., Giral, P., Elghouzzi, M. H., Jullien, A. M. & Poupon, R. 1993. Risk factors for hepatitis C virus infection in hepatitis C virus antibody ELISA-positive blood donors according to RIBA-2 status: a case-control survey. *Hepatology,* 17**,** 183-7.

Sgourakis, G., Gockel, I., Lyros, O., Lanitis, S., Dedemadi, G., Polotzek, U., Karaliotas, C. & Lang, H. 2012. The Use of Neural Networks in Identifying Risk Factors for Lymph Node Metastasis and Recommending Management of T1b Esophageal Cancer. *The American Surgeon,* 78**,** 195-206.

Shaheryar, Z. A. 2012. *Prevalence of hepatitis C* [Online]. Lahore. Available: http://dawn.com/2012/04/19/prevalence-of-hepatitis-c/ [Accessed 3 July 2012].

Shahzad, F., Atiq, M., Ejaz, S. & Hameed, S. 2001. Hepatitis E: Review of a disease endemic in Pakistan. *JOURNAL-PAKISTAN MEDICAL ASSOCIATION,* 51**,** 166-169.

Shaikh, F. H., Ali Abro, H., Ali Chhutto, M., Abbasi, P. A., Shaikh, A. W. & Ali Buriro, S. 2009. Hepatitis C: frequency and risk factors associated with sero-positivity among adults in Larkana City. *J Ayub Med Coll Abbottabad,* 21**,** 107-9.

Sharma, S. D. 2009. Why No Vaccine For Hepatitis C Virus Yet? *JK Science: Journal of Medical Education & Research,* 11**,** 99-101.

Shazi, L. & Abbas, Z. 2006. Comparison of risk factors for hepatitis B and C in patients visiting a gastroenterology clinic. *J Coll Physicians Surg Pak,* 16**,** 104-7.

Shepard, C. W., Finelli, L. & Alter, M. J. 2005. Global epidemiology of hepatitis C virus infection. *The Lancet infectious diseases,* 5**,** 558-567.

Sherriff, A. & Ott, J. 2004. Artificial neural networks as statistical tools in epidemiological studies: analysis of risk factors for early infant wheeze. *Paediatric and perinatal epidemiology,* 18**,** 456-463.

Sherriff, L. C. & Mayon-White, R. 2003. A survey of hepatitis C prevalence amongst the homeless community of Oxford. *Journal of Public Health,* 25**,** 358-361.

Shi, H. Y., Lee, K. T., Wang, J. J., Sun, D. P., Lee, H. H. & Chiu, C. C. 2012. Artificial Neural Network Model for Predicting 5-Year Mortality After Surgery for Hepatocellular Carcinoma: A Nationwide Study. *Journal of Gastrointestinal Surgery***,** 1-6.

Shin, H. R., Kim, J. Y., Ohno, T., Cao, K., Mizokami, M., Risch, H. & Kim, S. R. 2000. Prevalence and risk factors of hepatitis C virus infection among Koreans in rural area of Korea. *Hepatol Res,* 17**,** 185-196.

Sievert, W., Altraif, I., Razavi, H. A., Abdo, A., Ahmed, E. A., Alomair, A., Amarapurkar, D., Chen, C. H., Dou, X. & El Khayat, H. 2011. A systematic review of hepatitis C virus epidemiology in Asia, Australia and Egypt. *Liver International,* 31**,** 61-80.

Slinker, B. K. & Glantz, S. A. 2008. Multiple Linear Regression Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable. *Circulation,* 117**,** 1732-1737.

Sonia, S. 2010. Little history of DHQ hospital Gujranwala. Available: http://dhqbrief.blogspot.com/ [Accessed 25th December 2013].

Southern, W. N., Drainoni, M. L., Smith, B. D., Christiansen, C. L., Mckee, D., Gifford, A. L., Weinbaum, C. M., Thompson, D., Koppelman, E. & Maher, S. 2011. Hepatitis C testing practices and prevalence in a high-risk urban ambulatory care setting. *Journal of viral hepatitis,* 18**,** 474-481.

Souto, F. J., Fontes, C. J., Pignati, L. T., Pagliarini, M. E., Menezes Vda, M., Martinelli Ade, L., Figueiredo, J. F., Donadi, E. A. & Passos, A. D. 2012. Risk factors for hepatitis C virus infection in Inland Brazil: an analysis of pooled epidemiological sectional studies. *J Med Virol,* 84**,** 756-62.

Su, Y. Y. & Wang, N. 2011. [Primary risk factors of hepatitis C virus infection: a Meta-analysis]. *Zhonghua Liu Xing Bing Xue Za Zhi,* 32**,** 940-5.

Sun, C. A., Chen, H. C., Lu, C. F., You, S. L., Mau, Y. C., Ho, M. S., Lin, S. H. & Chen, C. J. 1999. Transmission of hepatitis C virus in Taiwan: prevalence and risk factors based on a nationwide survey. *J Med Virol,* 59**,** 290-6.

Sun, Y., Meng, Z., Wang, S., Chen, X., Sun, D., Chen, Z., Liu, C. & Zhuang, H. 1991. Epidemiologic investigation on an outbreak of hepatitis C. *Chinese medical journal,* 104**,** 975-979.

Tabaton, M., Odetti, P., Cammarata, S., Borghi, R., Monacelli, F., Caltagirone, C., Bossù, P., Buscema, M. & Grossi, E. 2010. Artificial neural networks identify the predictive values of risk factors on the conversion of amnestic mild cognitive impairment. *Journal of Alzheimer's Disease,* 19**,** 1035-1040.

Talpur, A. A., Memon, N., Solangi, R. & Ghumro, A. 2007. Knowledge and attitude of patients towards hepatitis B and C. *Pak J surg,* 23**,** 162-165.

Tanwandee, T., Piratvisuth, T., Phornphutkul, K., Mairiang, P., Permpikul, P. & Poovorawan, Y. 2006. Risk factors of hepatitis C virus infection in blood donors in Thailand: a multicenter case-control study. *J Med Assoc Thai,* 89 Suppl 5**,** S79-83.

Taylor, R., Taylor, A. & Smyth, J. 2002. Using an artificial neural network to predict healing times and risk factors for venous leg ulcers. *Journal of wound care,* 11**,** 101-105.

Temkin, N. R., Holubkov, R., Machamer, J. E., Winn, H. R. & Dikmen, S. S. 1995. Classification and regression trees (CART) for prediction of function at 1 year following head trauma. *Journal of neurosurgery,* 82**,** 764-771.

Thaikruea, L., Thongsawat, S., Maneekarn, N., Netski, D., Thomas, D. L. & Nelson, K. E. 2004. Risk factors for hepatitis C virus infection among blood donors in northern Thailand. *Transfusion,* 44**,** 1433-40.

Thang, N. D., Erhart, A., Speybroeck, N., Hung, L. X., Thuan, L. K., Hung, C. T., Ky, P. V., Coosemans, M. & D'alessandro, U. 2008. Malaria in central Vietnam: analysis of risk factors by multivariate analysis and classification tree models. *Malaria Journal,* 7**,** 28.

Thimme, R., Oldach, D., Chang, K.-M., Steiger, C., Ray, S. C. & Chisari, F. V. 2001. Determinants of viral clearance and persistence during acute hepatitis C virus infection. *The Journal of experimental medicine,* 194**,** 1395-1406.

Tong, C., Khan, R., Beechng, N., Tariq, W., Hart, C., Ahmad, N. & Malik, I. 1996. The occurence of hepatitis B and C viruses in Pakistani patients with chronic liver disease and hepatocellular carcinoma. *Epidemiology and infection,* 117**,** 327-332.

Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology,* 49**,** 1225-1231.

Tufféry, S. 2011. *Data Mining and Statistics for Decision Making*, Wiley.

Ture, M., Kurt, I., Turhan Kurum, A. & Ozdamar, K. 2005. Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications,* 29**,** 583-588.

Umar, M. & Bilal, M. 2012. Hepatitis C, A Mega Menace: A Pakistani Perspective. *J Pak Med Stud,* 2**,** 68-72.

Umar, M. & Hamama-Ul-Bushra. 2006. *Hepatitis C in Pakistan,* Rawalpindi, SAF.

Umar, M., Tul Bushra, H., Ahmad, M., Khurram, M., Usman, S., Arif, M., Adam, T., Minhas, Z., Arif, A. & Naeem, A. 2010. Hepatitis C in Pakistan: a review of available data. *Hepatitis Monthly,* 10**,** 205.

Van Beek, I., Buckley, R., Stewart, M., Macdonald, M. & Kaldor, J. 1994. Risk factors for hepatitis C virus infection among injecting drug users in Sydney. *Genitourin Med,* 70**,** 321-4.

Van Den Hoek, J. A., Van Haastrecht, H. J., Goudsmit, J., De Wolf, F. & Coutinho, R. A. 1990. Prevalence, incidence, and risk factors of hepatitis C virus infection among drug users in Amsterdam. *J Infect Dis,* 162**,** 823-6.

Van Der Poel, C., Cuypers, H., Reesink, H., Choo, Q. L., Kuo, G., Han, J., Quan, S., Polito, A., Verstraten, J., Van De Wouw, J. & Et Al. 1991. Risk factors in hepatitis C virus-infected blood donors. *Transfusion,* 31**,** 777-9.

Vineis, P., Rainoldi, A. & Tu, J. 1997. NEURAL NETWORKS AND LOGISTIC REGRESSION: ANALYSIS OF A CASE-CONTROL STUDYON MYOCARDIAL INFARCTION. AUTHOR'S REPLY. *Journal of clinical epidemiology,* 50**,** 1309-1310.

Voss, R., Cullen, P., Schulte, H. & Assmann, G. 2002. Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks. *International journal of epidemiology,* 31**,** 1253-1262.

W.H.O. 2003. *"Hepatitis C"* [Online]. Available: http://www.who.int/csr/disease/hepatitis/whocdscsrlyo2003/en/index4.html [Accessed 6 July 2012].

Waheed, Y., Shafi, T., Safi, S. Z. & Qadri, I. 2009. Hepatitis C virus in Pakistan: A systematic review of prevalence, genotypes and risk factors. *World journal of gastroenterology: WJG,* 15**,** 5647.

Wang, D., Wang, Q., Shan, F., Liu, B. & Lu, C. 2010. Identification of the risk for liver fibrosis on CHB patients using an artificial neural network based on routine and serum markers. *BMC infectious diseases,* 10**,** 251.

Wang, L.-J., Lin, S.-K., Chiang, S.-C., Su, L.-W. & Chen, C.-K. 2013. Risk Factors for HIV, Viral Hepatitis, and Syphilis among Heroin Users in Northern Taiwan. *Substance use & misuse,* 48**,** 89-98.

Wazir, M. S., Mehmood, S., Ahmed, A. & Jadoon, H. R. 2008. Awareness among barbers about health hazards associated with their profession. *J Ayub Med Coll Abbottabad,* 20**,** 35-8.

West, P. M., Brockett, P. L. & Golden, L. L. 1997. A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science***,** 370-391.

Wie, X. W., Zhou Bs 2007. Determining the risk factors of uterine myomas by using back propagation neural network. *Zhonghua Yu Fang Yi Xue Za Zhi,* 1.

Wing, E. A. S. 2011. *Highlights: Pakistan Economic Survey 2011-12* [Online]. Islamabad. Available: http://www.finance.gov.pk/survey/chapter_12/highlights.pdf [Accessed].

Wise, M., Finelli, L. & Sorvillo, F. 2010. Prognostic factors associated with hepatitis C disease: a case-control study utilizing US multiple-cause-of-death data. *Public Health Reports,* 125**,** 414.

Wolff, F. H., Fuchs, S. C., Barcellos, N. T., Falavigna, M., Cohen, M., Brandao, A. B. & Fuchs, F. D. 2008. Risk factors for hepatitis C virus infection in individuals infected with the HIV. *Dig Liver Dis,* 40**,** 460-7.

Wong, W., Fos, P. J. & Petry, F. E. 2003. Combining the performance strengths of the logistic regression and neural network models: a medical outcomes approach. *Scientific World Journal,* 3**,** 455-476.

Xue, H., Tatsumi, N., Park, K., Shimizu, M., Kyojima, T., Sumiya, Y., Kawabata, S., Maeda, N. & Sakano, D. 1996. Searching for risk factors using multilayer neural network as a classifier. *Informatics for Health and Social Care,* 21**,** 229-232.

Yabuuchi, T., Tsukuma, H., Hiyama, T. & Nakanishi, K. 1993. Risk factors of hepatitis C virus infection]. *[Nihon kōshū eisei zasshi] Japanese journal of public health,* 40**,** 1006.

Yazdanpanah, Y., De Carli, G., Migueres, B., Lot, F., Campins, M., Colombo, C., Thomas, T., Deuffic-Burban, S., Prevot, M. H., Domart, M., Tarantola, A., Abiteboul, D., Deny, P., Pol, S., Desenclos, J. C., Puro, V. & Bouvet, E. 2005. Risk factors for hepatitis C virus transmission to health care workers after occupational exposure: a European case-control study. *Clin Infect Dis,* 41**,** 1423-30.

Yin, W., Tan, C., Chen, M., Liu, Z. & Wang, Z. 2004. Management of Disposable Infusion Set and Injection Syringe [J]. *Chinese Journal of Nosoconmiology,* 2.

Zakizad, M., Salmeh, F., Yaghoobi, T., Yaghoubian, M., Nesami, M. B., Esmaeeli, Z., Vaezzadeh, N., Shahmohammadi, S., Modanloo, S., Sadeghian, A. A., Abdolmanafi, S. J., Mohammadpour, R. A., Siamian, H. & Khosravi, A. 2009. Seroprevalence of hepatitis C infection and associated risk factors among addicted prisoners in Sari-Iran. *Pak J Biol Sci,* 12**,** 1012-8.

Zaller, N., Nelson, K. E., Aladashvili, M., Badridze, N., Del Rio, C. & Tsertsvadze, T. 2004. Risk factors for hepatitis C virus infection among blood donors in Georgia. *Eur J Epidemiol,* 19**,** 547-53.

Zeuzem, S., Teuber, G., Lee, J. H., Ruster, B. & Roth, W. K. 1996. Risk factors for the transmission of hepatitis C. *J Hepatol,* 24**,** 3-10.

Zhao, Y., Shen, L., Ma, J., Gao, Z., Han, X., Qi, S. & Li, Q. 2013. Epidemiology of Hepatitis C Virus Infection and Risk Factor Analysis in the Hebei Province, China. *PLoS One,* 8**,** e75586.

Zibran, M. F. 2007. CHI-Squared Test of Independence. Department of Computer Science, University of Calgary, Alberta, Canada.[online].[Cited 2010-08-12]. Available on the internet:< http://pages. cpsc. ucalgary. ca/~ saul/wiki/uploads/CPSC681/topic-fahim-CHI-Square. pdf.

Zurada, J. & Lonial, S. 2011. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. *Journal of Applied Business Research (JABR),* 21.

Appendix 1

QUESTIONNAIRE

Outcome	**Case/Control**
Date of visit:	Hospital Name:
Registration No.	Patient of Name
Address:	
District:	

Socio-Demographic Characteristics	
Gender:	Male ❑ Female ❑
Marital status	Ever-married ❑ Never-married ❑
Age (in years)	
Family Status	Single ❑ Nuclear ❑
Residential Area	Urban ❑ Rural ❑
Monthly household income (in PKR)	
Family size	
House status	Owned ❑ Rented ❑
No. of persons sharing the room	
Health insurance	Yes ❑ No ❑
Patients' Education (in years)	
Father education	Literate ❑ illiterate ❑
Mother education	Literate ❑ illiterate ❑

Risk Factors	
i. Medical / Health Care Risk Factors	
Patient past history of jaundice	Yes ❑ No ❑
Patients history of liver disease	Yes ❑ No ❑
History of blood transfusion	Yes ❑ No ❑
History of blood donation	Yes ❑ No ❑
Have you currently or in past undergone dental surgery	Yes ❑ No ❑
Have you had kidney dialysis?	Yes ❑ No ❑
History of accidental needle click.	Yes ❑ No ❑
History of therapeutic injections	Yes ❑ No ❑
Major/Minor surgery	Yes ❑ No ❑
History of Angiography/Angioplasty	Yes ❑ No ❑
History of endoscopy/gastroscopy	Yes ❑ No ❑
History of hospitalization	Yes ❑ No ❑
History of Colonoscopy	Yes ❑ No ❑
Road Traffic Accident	Yes ❑ No ❑
History of Cuts (please specify)	Yes ❑ No ❑

History of branulla insertion	Yes ❑ No ❑
History of Acupuncture	Yes ❑ No ❑
Have you had eye surgery?	Yes ❑ No ❑
Ear wax removal from hospital	Yes ❑ No ❑
Migration/travelling	Yes ❑ No ❑
Cesarean Section (Female only)	Yes ❑ No ❑
Place of Cesarean Section (Female only)	None ❑ Govt. Hospital ❑ Private Hospital ❑ Midwife/LHV❑
History of Abortion/D&C (Female only)	Yes ❑ No ❑
ii. Behavioral Characteristics	
Habit of tattooing	Yes ❑ No ❑
Sharing of Razors	Never ❑ Rarely ❑ Often ❑
Facial barber shave	Yes ❑ No ❑
Habit of nail cutting from barber shop	Yes ❑ No ❑
Under shave from barber "Hammam"	Yes ❑ No ❑
Armpit shave by barber	Yes ❑ No ❑
Extramarital relationship	No ❑ Yes ❑ Non-response ❑
Had ear/nose piercing?	Yes ❑ No ❑
Have you more than one marriage?	Yes ❑ No ❑
Have you injected drugs, even once?	Yes ❑ No ❑
Sharing of syringes	Yes ❑ No ❑
Habit of sharing of toothbrush/Miswak	Yes ❑ No ❑
Sharing of nail cutter	Yes ❑ No ❑
Ever imprisoned in past?	Yes ❑ No ❑
Sharing of eye connectors or lenses	Yes ❑ No ❑
Minor surgery by barber	Yes ❑ No ❑
Homelessness and hostel life in past	Yes ❑ No ❑
Alcohol use	Yes ❑ No ❑
iii. Family History	
Family history of hepatitis	Yes ❑ No ❑
Family history of jaundice/liver disease	Yes ❑ No ❑

Appendix 2: Summary Table of Literature Review

Author	Study site	Study Design	Sample size	Significant risk factors found
Alter et al (1990)	United States	Descriptive	310	Intravenous drug use (IDUs)
Girardi et al., (1990)	Italy	Descriptive	80	Sharing of needles
Maisonneuve et al (1991)	France	Descriptive	117	blood transfusion, IDUs and gayorientation
An-der-Poel et al., 1991	Netherland	Descriptive	125	blood transfusion, IDUs and gay orientation
Kolho and Krusius (1992)	Finland	Descriptive	305	Low level of education, parenteral source of infection
Pohjanpelto (1992)	Finland	Descriptive	160	Exposure with blood and history of travelling.
Kaldor et al (1992)	Sydney	Case control	Cases=430 Controls=210	history of injecting drug use (IDUs)
Darwish et al (1993)	Egypt	Case control	160	Older age, H/O injections and H/O schistosomiasis
Kaldor et al (1992)	France	Case Control	Cases=220 Controls=210	history of jaundice, blood transfusion and injecting drug users (IDUs)
Neal et al (1994)	UK	Case Control	Cases=300 Controls=150	tattooing, ear piercing, travelling, acupuncture
Montagnac et al (1994)	Brazil	Case Control	138	Gender (male), older age and nonwhite ethnicity
Mele et al (1994)	Italy	Case Control	Cases=342 controls=1095	surgical interventions, dental surgery, hospitalization, other percutaneous exposures and multiple marital relation
Chiaramonte et al (1996)	Italy	Case Control	Cases=500 controls=500	blood transfusion and drug addiction, of non-disposable needles, hospitalization
Kim et al (1996)	Korea	Case Control	Cases=64 controls=128	H/O acute hepatitis and transfusion
Mohamed et al (1996)	Egypt	Cross Sectional	5071	age>25, male, being married, rural residence, injections
Comandini et al (1998)	Rome, Italy	Case Control	Cases=365 controls=465	intravenous injections, minor surgical procedures, and blood transfusion
Balasekaran et al (1999)		Case Control	Cases=477 controls=477	Heavy Alcohol Intake, marital relationship, tattoo

Delage et al (1999)	Canada	Case Control	Cases=267 controls=1068	history of imprison, intravenous drug use, blood transfusion, tattooing and marital relation
Merle et al (1999)	France	Case Control	Cases= controls=	medical procedure after an accident and family history
Sun et al., 1999	Taiwan	Case Control	Cases=272 controls=282	Transfusion, therapeutic injections, and surgical procedure
Briggs et al (2001)	Urban Veterans	Case Control	Cases=185 controls=847	IDUs, Tattooing, and transfusion
Habib et al (2001)	Egypt	Cross sectional	3993	male gender, marriage, blood transfusion, anti-schistosomiasis injection, cesarean section or abortion in (females only)
Alavian et al (2002)	Iran	Case Control	Cases=193 controls=196	Transfusion, extramarital marital relationship, non-intravenous drug, endoscopy, and receiving wounds at war
(Brandao and Fuchs, 2002	Brazil	Case Control	Cases=178 Controls=356	marital relationship with a positive partner
Kim et al (2002)	Korea	Case Control	Cases=178 controls=226	Endoscopy, blood transfusion
Mishra et al (2003)	North Florida and South Georgia	Cross Sectional	274	use of illicit drugs, incarceration, Low income
Thaikruea et al., 2004	Thailand	Case Control	Cases=166 controls=166	multiple marital relation
Hand and Vasquez (2005)	Texas–Mexico border	Case Control	Cases=320 controls=307	(IDUs), tattooing, blood transfusion
Yazdanpanah et al (2005)	Europe	Case Control	Cases=60 controls=204	Occupational transmission through needle stick and deep injury
Hajiani et al., 2006a	Iran	Case Control	Cases=254 controls=260	Tattooing, blood transfusion
Karmochkine et al (2006a)	French	Case Control	Cases=500 controls=750	wound care, beauty treatment etc.
Ben-Alaya Bouafif et al (2007)	South Africa	Case Control	Cases=57 controls=285	history of invasive procedures etc.
Nguyen et al (2007)	Vietnam	Case Control	837	Hospitalization and Tattooing

Liu et al (2009)	China	Case Control	Cases=69 controls=207	unregulated medical procedures
Awadalla et al (2011)	Egypt	Cross Sectional	Cases=168 controls=832	poor socio-economic status, being married, Accidental wound
He et al (2011)	China	Case Control	Cases=305 controls=610	razor sharing, acupuncture, hospitalization, family history of hepatitis
Kandeel et al., (2012)	Egypt	Case Control	Cases=86 controls=287	reuse of syringes, imprisonment, IV fluid in hospital, minor surgical operations, hospitalization
Luby et al (1997)	Hafizabad District of Punjab	Case Control	Cases=15 controls=67	tattooing, injecting drug users, barber shave, sharing razor, ear/nose piercing, sharing tooth
Bari et al (2001)	Rawalpindi-Islamabad	Case Control	Cases=57 controls=180	daily shave and armpit shave by a barber
Janjua and Nizamy (2004)	Rawalpindi-Islamadad	cross sectional		reusing of razor
Akhtar et al., 2004	Karachi	Case Control	Cases=80 controls=160	hospitalization, theraputic injections and multiple marital relation
Shazi and Abbas (2006)	Karachi	Case Control	Cases=63 controls=44	low education status, blood transfusions, occupational exposure, barber shave
Ijaz and Akhter (2007)	Punjab	Case Control	Cases=135 controls=195	surgery, barber shave and blood transfusion
Abbas et al (2008)	Rural Sindh	Cross Sectional	843	dental procedures, history of hepatitis
Ghaffar et al (2009)	Quetta, Balochistan	Case Control	Cases=108 controls=108	history of injections, family history of jaundice and previous surgeries
Qureshi et al (2009)	Karachi	Cross Sectional	1446	dental surgery, blood transfusion, I/V and I/M history of injections, family history of hepatitis, surgery